生物信息学札记（第3版）

樊龙江

浙江大学作物科学研究所
浙江大学生物信息学研究所
浙江大学IBM生物计算实验室
浙江大学沃森基因组科学研究院

2010年1月

本材料可通过下列网址获得:
http://ibi.zju.edu.cn/bioinplant/

前言

第一版
这份材料是我学习和讲授《生物信息学》课程时的备课笔记，材料大多是根据当时收集的一些外文资料翻译编辑而成。学生在学习过程中经常要求我给他们提供一些中文的讲义或材料，这促使我把我的这份笔记整理并放到网上，供大家参考。要提醒使用者的是，这份材料仅是根据我对生物信息学的一些浮浅的认识整理而成，其中的错误和偏颇只能请读者自鉴了。

2001年6月

第二版
自1999年开始接触生物信息学以来，一晃已近六年，而本札记也近四岁了。2001和2002年中国科学院理论物理所的郝柏林院士在浙江大学首次开设生物信息学研究生课程，我作为他的助教系统地学习了生物信息学；同时，借着我国水稻基因组测序计划的机遇，在他的带领下从2001年开始从事水稻基因组分析，从此自己便完全投入到这一崭新、引人入胜的领域中来。
不断有来信向我索要本札记的电子版文件，同时在不少网站上看到推荐该札记的内容。生物信息学、基因组学等发展很快，现在再回头审看该札记，有些部分已惨不忍读，这促使我下

决心更新它。但因时间和学识问题，还是有不少部分自己不甚满意，就只有待日后再努力了。我的硕士生温晓协助收集了部分资料。欢迎告诉我札记中的BUG，我的信箱fanlj@zju.edu.cn 或bioinplant@zju.edu.cn。

2005年3月30日

第三版

近年来高通量测序技术产生的序列数据大量出现（如小RNA和大规模群体SNP数据），本次更新根据这一进展增加了两章内容，分别是第七章有关小RNA的分析和第八章遗传多态性及正向选择检测。两章内容由我的博士生王煜为主编写，李泽峰和刘云参与了文献整理。另外还更新了第四章有关水稻基因组分析一节。

2010年1月

简要目录

第一章 生物信息学通论

第一节 生物信息与生物信息学

一、迅速膨胀的生物信息

二、生物信息学的概念

第二节 生物信息学发展简史

第三节 基因组时代：生物信息学的应用与展望

第二章 分子数据库

第一节 初级数据库

一、DNA数据库

二、基因组数据库

三、蛋白质序列数据库

四、蛋白质结构数据库

第二节 初级序列数据的注释

第三节 数据库信息检索系统

第四节 数据库的冗余与偏误

第五节 向数据库发送序列数据及其它

第三章 序列分析与比较

第一节 序列组成和单一序列分析

一、碱基组成

二、碱基相邻频率

三、同向重复序列分析

四、DNA序列的几何学分析——Z曲线

第二节 序列联配

一、Needleman-Wunsch算法

二、Smith-Waterman算法

三、序列相似性统计特征

1、二进制值或标准比值（Bit Score）；2、P值（P-value）；3、BLAST和FASTA的数据库搜索策略；4、空位列线（gapped alignment）的统计问题；5、边际效应（edge effect）；6、替换矩阵的选择；7、空位罚值（gap penalties）

四、替换矩阵

1、替换矩阵的一般原理；2、PAM氨基酸替换矩阵；3、BLOSUM氨基酸替换矩阵；4、DNA替换矩阵

五、多序列联配

第三节 数据库搜索引擎——BLAST和FASTA应用

一、数据之海与一叶轻舟

二、BLAST：核酸数据库搜索

1、BLAST实战操作（1）；2、BLAST的检索报告；3、BLAST选项；4、BLAST实战操作（2）

三、BLAST：蛋白质数据库搜索

四、FASTA：另一种搜索策略

1、FASTA选项；2、FASTA实战操作及其检索报告

第四节 寡核苷酸设计

一、寡核苷酸设计

1、引物设计；2、用于检测相关基因的简并探针

第四章 基因组测序与分析

第一节 DNA测序与序列片段的拼接

一、DNA测序的一般方法

1、DNA测序的基本原理；2、双脱氧测序法（Sanger法）；3、化学测序法（Maxam-Gilbert法）；4、荧光自动测序仪

二、DNA片段测序策略

1、从遗传图谱、物理图谱到基因组序列图谱；2、鸟枪测序法（shotgun sequencing）；3、引物步查法（primer walking）；4、限制性酶切-亚克隆法（restriction endonuclease digestion and subcloning）

三、基因组测序策略

四、序列片段的拼接方法

五、EST测序

第二节 基因组注释：基因区域的预测

一、从序列中寻找基因

1、基因及基因区域预测；2、发现基因的一般过程；3、解读序列（making sense of the sequence）

二、最长ORF法等：基于编码区特性

三、序列相似性比较法

四、隐马尔可夫模型（HMM）

五、神经网络

六、RNA二级结构预测

第三节 基因组分析

一、基因组分析：生物信息学发展的“史记”

二、比较基因组学

第四节 基因组分析举例：水稻基因组分析

一、 现代的二倍体，古老的多倍体

二、 最小的核基因组：基因组在扩增还是在缩小?

三、 籼粳稻分化时间比原来估计的要迟得多

四、 水稻高GC含量基因的进化机制

五、 水稻小RNA可能是驯化和育种选择的靶基因

第五章 分子进化

第一节 系统树及其它

附录：

第一章 生物信息学通论

我们处在一个激动人心的时代——基因组时代。科学的进步已使人类可以窥探生命的秘密，甚至包括人类自身。人类基因组在世纪之交被人类自己破译了。这部由 30 亿个字符组成的人类遗传密码本已活生生地摆在了我们面前。于此同时，来自其它生物的基因组信息源源不断从自动测序仪中涌出，堆集如山，浩如烟海。这些海量的生物信息是用特殊的“遗传语言”——DNA 的四个碱基字符(A、T、G 和 C)和蛋白质的 20 个氨基酸字符(A、R、N、D、C、Q、E、G、H、I、L、K、M、F、P、S、T、W、Y 和 V)——写成。

《科学》（*Science*）在 2001 年 2 月 16 日人类基因组专刊上配发了一篇题为“生物信息学：努力在数据的海洋里畅游”(Roos DS. Bioinformatics—Trying to swin in a sea of data. Science，2001，291：1260-1261)的文章。文章写道：“我们身处急速上涨的数据海洋中…，我们如何避免生物信息的没顶之灾呢？”一叶轻舟也许可以救命！生物信息学便是我们找到的这样一条“轻舟”，而且我们已在这条轻舟上安装了诸如卫星定位系统等先进的电子设备。也许在不久的将来，人类会造就一艘永不沉没的航空母艇……生物信息学是一门年青的学科，学科虽然年青，但它充满挑战、机遇且引人入胜。

第一节 生物信息与生物信息学

一、迅速膨胀的生物信息

近 20 年来，分子生物学发展的一个显著特点是生物信息的剧烈膨胀，且迅速形成了巨量的生物信息库。这里所指的生物信息包括多种数据类型，如分子序列(核酸和蛋白质)，蛋白质二级结构和三维结构数据、蛋白质疏水性数据等等。由实验获得的大量核酸序列和三维结构数据被存在数据库中，这些数据库就是所谓的初级数据库(primary databases)；那些由原始数据分析而来的诸如二级结构、疏水位点和功能区(domain)数据，则组成了所谓的二级数据库(secondary databases)。那些由核酸数据库序列翻译而来的蛋白质序列数据组成的蛋白质数据库，也应被视为二级数据库。

生物信息的增长是惊人的。近年来，核酸库的数据每 10 个月左右就要翻一翻，2000 年底，数据库数据则达到了创记录的 100 亿个记录，大量生物(甚至包括我们人类自身)的整个基因组序列被测定完成或正在进行中，遍布世界各地研究实验室的高通量大型测序仪在日夜不停地运转，每天都有成千上万的数据被源源不断地输入相应的生物信息库中。同时，由这些原始数据分析加工而来的蛋白质结构等数据信息也被世界各地的分子生物学、生物信息学等学科领域专家输入二级数据库中。图 1.1 显示出了各种生物信息的同步增长状况。

迅速膨胀的生物信息给科学家们提出了一个新问题：如何有效管理、准确解读、充分使用这些信息？

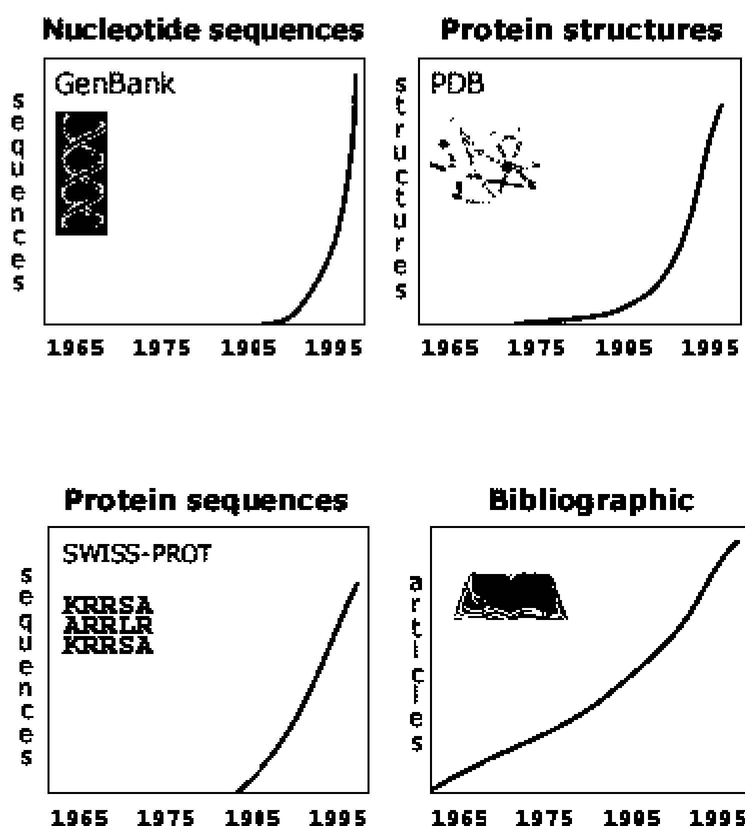


图 1.1 各类生物信息的同步增长状况。图中依次为核酸序列（GenBank)、蛋白质序列（PDB)、蛋白质序列（SWISS-PROT）和文献数量增长幅度（引自 NCBI，2000)。

二、生物信息学的概念

生物信息学便是在生物信息的急剧膨胀的压力下诞生了。

一般意义上，生物信息学是研究生物信息的采集、处理、存储、传播、分析和解释等各方面的一门学科，它通过综合利用生物学、计算机科学和信息技术而揭示大量而复杂的生物数据所赋有的生物学奥秘。具体而言，生物信息学作为一门新的学科领域，它是把基因组 DNA 序列信息分析作为源头，在获得蛋白质编码区的信息后进行蛋白质空间结构模拟和预测，然后依据特定蛋白质的功能进行必要的药物设计。基因组信息学、蛋白质空间结构模拟以及药物设计构成了生物信息学的 3 个重要组成部分。从生物信息学研究的具体内容上看，生物信息学应包括这 3 个主要部分：(1)新算法和统计学方法研究；(2)各类数据的分析和解释；(3)研制有效利用和管理数据新工具。Claverie （2000）的一段英文描述如下：“Bioinformatics is the science of using information to understand biology. It’s the discipline of obtaining information about genomic or protein sequence data. This may involve similarity searches of databases, comparing your unidentified sequence to the sequences in a database, or making predictions about the sequence based on current knowledge of similar sequences.”

生物信息学最初更多地是关注数据库，那些数据库存储着来自基因组测序计划完成的序列数据。目前生物信息学已今非昔比，它所关注的是各类数据，包括生物大分子的三维结构、代谢途径和基因表达等等。生物信息学最使人们感兴趣的是它利用计

算方法分析生物数据，如根据核酸序列预测蛋白质序列、结构、功能的算法等。虽然这些预测还不是非常精准，但是当可靠的实验数据还无法得到的情况下，这这一预测可以作为一盏路灯，指示你应如何开展实验。

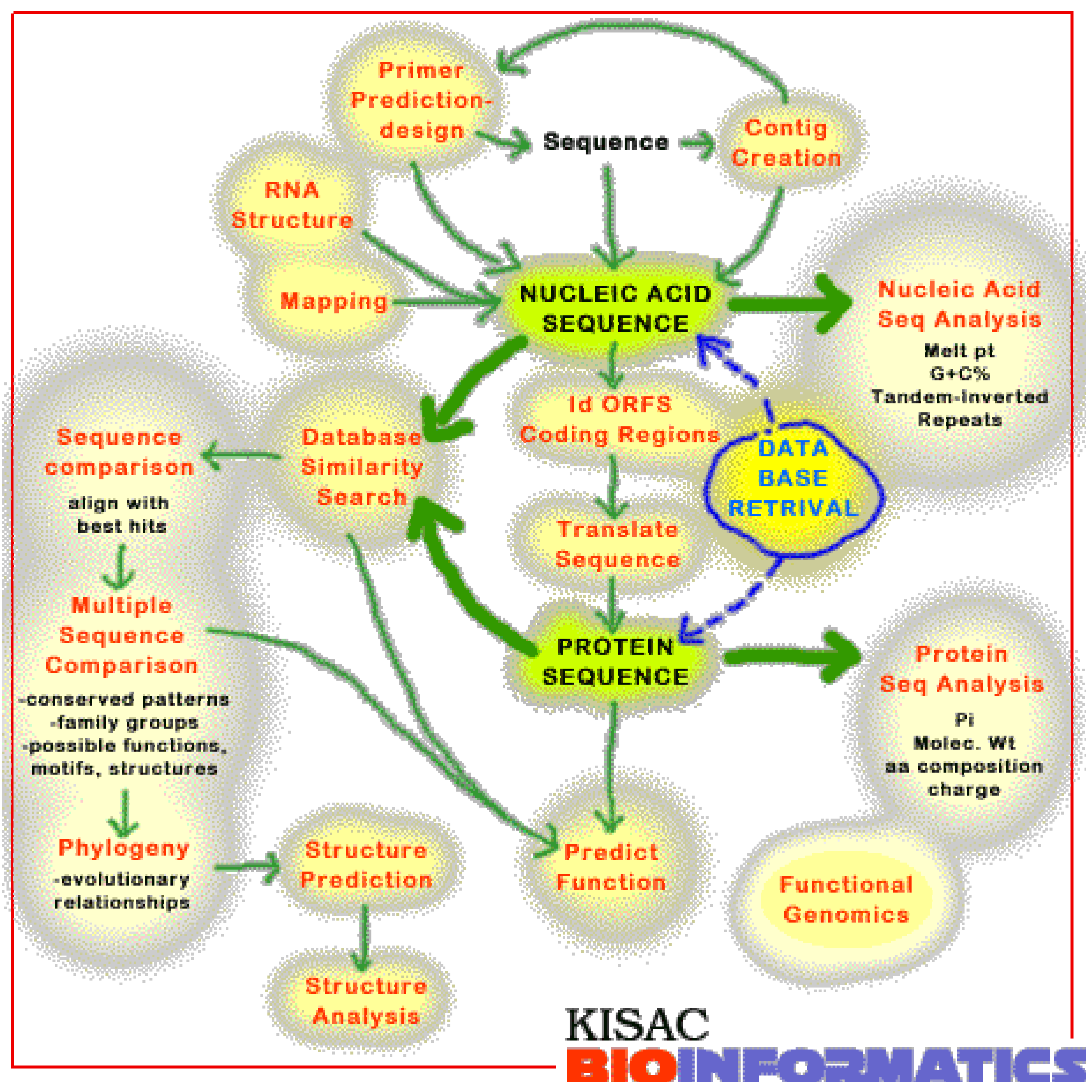


图 1-2 生物信息学“路线图”。取自http://www.kisac.ki.se/。

生物信息学的诞生和发展最早可以追溯到上个世纪的 60 年代，波林(Pauling)分子进化理论的出现，已预示着生物信息学的来临。而真正意义上的“生物信息学(Bioinformatics)”一词的出现则是 1990 年(见："A term coined in 1990 to define the use of computers in sequence analysis" (Claverie, 2000)，据说是由出生在马来西亚的美籍学者林华安（Hwa A. Lim）首次使用的（郝柏林和张淑誉，2002)。

虽然生物信息学的历史并不长，但正象生物信息的迅猛发展一样，生物信息学已发展了大量独具学科特色的分析方法和分析软件。例如，当获得了大量序列数据以后，我们现在已能进行序列家族或同源性分析；进行序列的聚类，建立进化树并确定序列间的进化关系；进行代谢途径相关基因的同源性分析，以及获取其它生物代谢途径的相关信息等。分析软件更是层出不穷，通过网络可以搜索到大量的相关信息。这些软件很多已成为商业化产品，但很多软件是可以免费获取的。这些分析软件(见附录)已

成为生物信息学最重要的研究手段，是生物学家获取信息的重要途径和生物信息学显示其价值的窗口。

NCBI Tools for data mining

PubMed Entrez BLAST OMIM Taxonomy Structure

Search GenBank for Go

NCBI
SITE MAP

BLAST
standard tool for sequence analysis

COGs
Clusters of Orthologous Groups

ORF finder
finds open reading frames

BLAST The Basic Local Alignment Search Tool (BLAST), for comparing gene and protein sequences against others in public databases, now comes in several flavors including PSI-BLAST, PHI-BLAST, and BLAST 2 sequences. Specialized BLASTs are also available for human, microbial, and malaria genomes, as well as for vector contamination, immunoglubulins, and tentative human consensus sequences.

Clusters of Orthologous Groups (COGs) currently covers 21 complete genomes from 17 major phylogenetic lineages. A COG is a cluster of very similar proteins found in at least three species. The presence or absence of a protein in different genomes can tell us about the evolution of the organisms, as well as point to new drug targets.

atg........tga ORF finder identifies all possible ORFs in a DNA sequence by locating the standard and alternative stop and start codons. The deduced amino acid

electronic PCR 001101011AGCGT Electronic PCR allows you to search your DNA sequence for sequence tagged site (STS), which have been used as landmarks in various types of genomic

图 1.3 美国国家生物技术信息中心（NCBI）网站数据分析工具网页。图中包括 BLAST、COG、ORF finder、Electronic PCR 等工具软件。

生物信息学还有另一个经常被使用的名字：“计算生物学”(computational biology)，此外“计算分子生物学”(computational molecular biology)和“生物分子信息学”(biomolecular informatics)等也被使用过。但严格意义上说，计算生物学的范围应更宽泛些[见“Strictly speaking, bioinformatics is a subset of the large field of computational biology, the application of quantitative analytical techniques in modeling biological system.” (Gibas and Jambeck, 2001)]。

正确认识和理解生物信息学这门新学科非常重要，它有助于该学科的科学研究和学习。《Bioinformatics》杂志的一篇社论文章(2000,vol 16 no.3，其翻译稿见庞洪泉和樊龙江，生物技术通报，2002，2：47-52)，评析了人们对生物信息学的一些不正确的认识：(1)“人人可以从事生物信息学研究”。这一认识的根源来自对生物信息学的 2 个误解，一是生物信息学研究不需大量经费投入，因为有如此多的数据资源，只要找本生物学教科书，有台电脑并连到国际网上，人人可以从事生物信息学研究；二是生物信息学的软件是免费的。殊不知生物信息的巨量特征目前向计算机提出了严峻的考验，而一台大型新型计算机可能要以千万甚至亿元计算，同时大量先进、最新的生物信息学分析软件包都是商业化产品，不付钱难以到；(2)“你最终还是需要具体的实验”。实验生物学家非常羡慕生物信息学家，认为“他们只是敲敲键盘，然后便是写论文”，他们的研究结果只是一种试验结果的预测，是对实验研究的一种“支持”。在分子生物学研究中，固定的模式应是先有某一假设，然后用某一实验去验证或支持

这一最初的猜测。在生物信息学研究中，也同样进行着这一模式：有一无效假设(例如某一序列在数据库中没有同源序列)，然后进行实验(如搜索数据库)并验证，拒绝或接受无效假设(如该序列的确有或无同源序列)。这是一个标准的假设—实验模式。在其它学科中，计算科学已被作为深入理解科学问题的重要手段，而在生物学领域还没有形成这样的共识；(3)“生物信息学是门新技术，但只是一门技术而已”。由此把生物信息学定位为一门新的应用学科。正如前面所说，虽然生物信息学是一门新学科，但在 60-70 年代，该学科最重要的一些算法便已被提出，生物计算和理论研究便形成雏形。把生物信息学仅仅作为一门应用技术，是从信息学移植来的技术应用于生物学科领域，这是一个致命的误解。生物信息学实际是一门充满丰富知识内涵的学科，它有很多尚待解决的科学问题。这些问题包括生物学方面的(如分子的功能如何进化)和计算方面的(如数据库系统间如何最有效地协同)。生物信息学不仅仅是一个技术平台，它同样需要周详的实验计划和准确的操作，同样需要丰富的想象和一瞬即逝的运气。

第二节 生物信息学发展简史

表 1.2 列出了生物信息学最近几十年的主要事件。这些事件大多是在“生物信息学”(bioinformatics)一词出现前便发生了。纵观生物信息学的发展历史，可将它分为 3 个主要阶段：(1)萌芽期(60-70 年代)：以 Dayhoff 的替换矩阵和 Neelleman-Wunsch 算法为代表，它们实际组成了生物信息学的一个最基本的内容和思路：序列比较。它们的出现，代表了生物信息学的诞生(虽然“生物信息学”一词很晚才出现)，以后的发展基本是在这 2 项内容上不断改善；(2)形成期(80 年代)：以分子数据库和 BLAST 等相似性搜索程序为代表。1982 年三大分子数据库的国际合作使数据共享成为可能，同时为了有效管理与日俱增的数据，以 BLAST、FASTA 等为代表工具软件和相应的新算法大量被提出和研制，极大地改善了人类管理和利用分子数据的能力。在这一阶段，生物信息学作为一个新兴学科已经形成，并确立了自身学科的特征和地位；(3)高速发展期(90 年代-至今)：以基因组测序与分析为代表。基因组计划，特别是人类基因组计划的实施，分子数据以亿计；基因组水平上的分析使生物信息学的优势得以充分表现，基因组信息学成为生物信息学中发展最快的学科前沿。Phred-Phrap-Consed 系统软件包自 1993 年出现，1995 年已广泛应用于鸟枪法测序中序列的碱基识别、拼装和编辑等，是目前人类基因组等测序计划的主要应用软件，与 BLAST 一起在人类基因组计划的研究历史中占有一席之地(见 *Science* 2001 年 2 月 16 日人类基因组专刊“A history of Human Genome Project”一文)。在此阶段，生物信息学已成为举世瞻目、竞相发展的热点学科。GenBank 等数据库中数据的增长在近十年来呈直线上升趋势(图 1.1)，这条曲线很容易就使我们联想到生物信息学的发展历程，可以说，这条曲线便是生物信息学近十余年发展的写照。生物信息学在近十余年间经历了长足的发展，并迅速成为生命科学新的生长点。人类基因组计划的实施和生物医药工业的介入是生物信息学迅猛发展的主要推动力。

英国剑桥大学出版社出版的《Bioinformatics》期刊(www.bioinformatics.oupjournal.org)是目前世界最知名生物信息学的学术期刊之一，它的前身是《Computer Applications in the Bioscience》(CABIOS)，1998 年更名为《Bioinformatics》。该杂志主要发表计算分子生物学、生物数据库和基因组生物信息学方面的文章。另外带有生物信息学字样的杂志还有《Applied Bioinformatics》、《Briefings in Bioinformatics》、《Journal of bioinformatics and computational biology》、《Genomics, proteomics & bioinformatics》、《Proceedings / IEEE

Computer Society Bioinformatics Conference》以及网上生物信息学杂志《BMC Bioinformatics》(www.biomedcentral.com)等。其它与生物信息学相关的出版物还很多，如《Nucleic Acids Research》、《Genome Research》、《Genomics》、《J. Mol. Biol.》、《BioTechniques》、《BioTechnology Software》等。

表 1.2　生物信息学发展的简史

1962	Pauling 提出分子进化理论
1967	Dayhoff 构建蛋白质序列数据库
1970	Needleman-Wunsch 算法被提出
1977	Staden 利用计算机软件分析 DNA 序列
1981	Smith-Waterman 算法出现
1981	序列模序(motif)的概念被提出(Doolittle)
1982	GenBank 数据库(Release3)公开；EMBL 创立
1982	λ-噬菌体基因组被测序
1983	Wilbur 和 Lipman 提出序列数据库的搜索算法(Wilber-Lipman 算法)
1985	快速序列相似性搜索程度 FASTP/FASTN 发布
1988	美国家生物技术信息中心(NCBI)创立
1988	欧洲分子生物学网络 EMBnet 创立；三大核酸数据库(GenBank、EMBL 和 DDBJ)开始国际合作
1990	快速序列相似性搜索程序 BLAST 发布
1991	表达序列标签(EST)概念被提出，从此开创 EST 测序
1993	英国 Sanger 中心在英国休斯顿建立
1994	欧洲生物信息学研究所在英国 Hinxton 成立
1995	第一个细菌基因组测序完成
1996	酶母基因组测序完成
1997	PSI-BLAST(BLAST 系列程序之一)发布
1998	PhilGreen 等人研制的自动测序组装系统 Phred-Phrap-Consed 系统正式发布
1998	多细胞线虫基因组测序完成
1999	果蝇基因组测序完成
2000	人类基因组测序基本完成
2001	人类基因组初步分析结果公布

*主要引自美国家生物信息中心(NCBI)Education-Bioinformatics Milestone(2000)，原文截止至 1999 年果蝇基因组测序完成，有关人类基因组、PhilGreen 等的自动测序组装系统和三大核酸数据库的合并等内容为作者补入。

**以上主要算法的原始文献出处：Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970 ，48(3):443-53；Staden R. Sequence data handling by computer. Nucleic Acids Res. 1977 ，4(11):4037-51；Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981，25;147(1):195-7；Doolittle RF. Similar amino acid sequences: chance or common ancestry? Science. 1981，214(4517):149-59；Wilbur WJ, Lipman DJ. Rapid similarity searches of nucleic acid and protein data banks. Proc Natl Acad Sci U S A. 1983，80(3):726-30；Lipman DJ, Pearson WR. Rapid and sensitive protein similarity searches. Science. 1985，227(4693):1435-41；karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc. Natl. Acad. Sci. USA, 1990, 87:2264-2268。

我们可从另一个角度来审视生物信息学的发展历程：美国家生物技术信息中心(NCBI)的十年(1989-1999)发展史，它是生物信息学近十余年来发展的一个缩影。NCBI的十年发展史(图 1.5)可以说是年年有进步：筹备(1989)、BLAST 启动(1990)、Entrez开始检索(1991)、GenBank 加盟(1992)、Entrez 上网和 3-D 分子数据建立(1993)、NCBI上网(1994)、解读序列(1995)、从序列中发基因(1996)、PubMed 登网和蛋白质分析(1997)、GenBank 碱基数据超过 10 亿(1998)、关注人类基因组(1999)。GenBank 十年来分子数据的增长曲线也正表明了 NCBI 的十年发展轨迹。

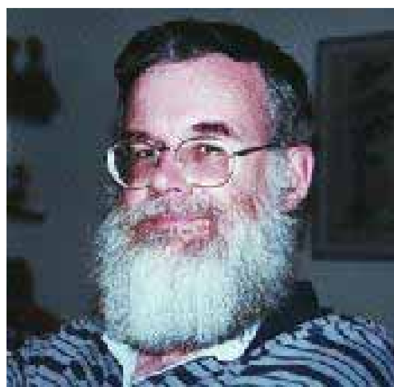
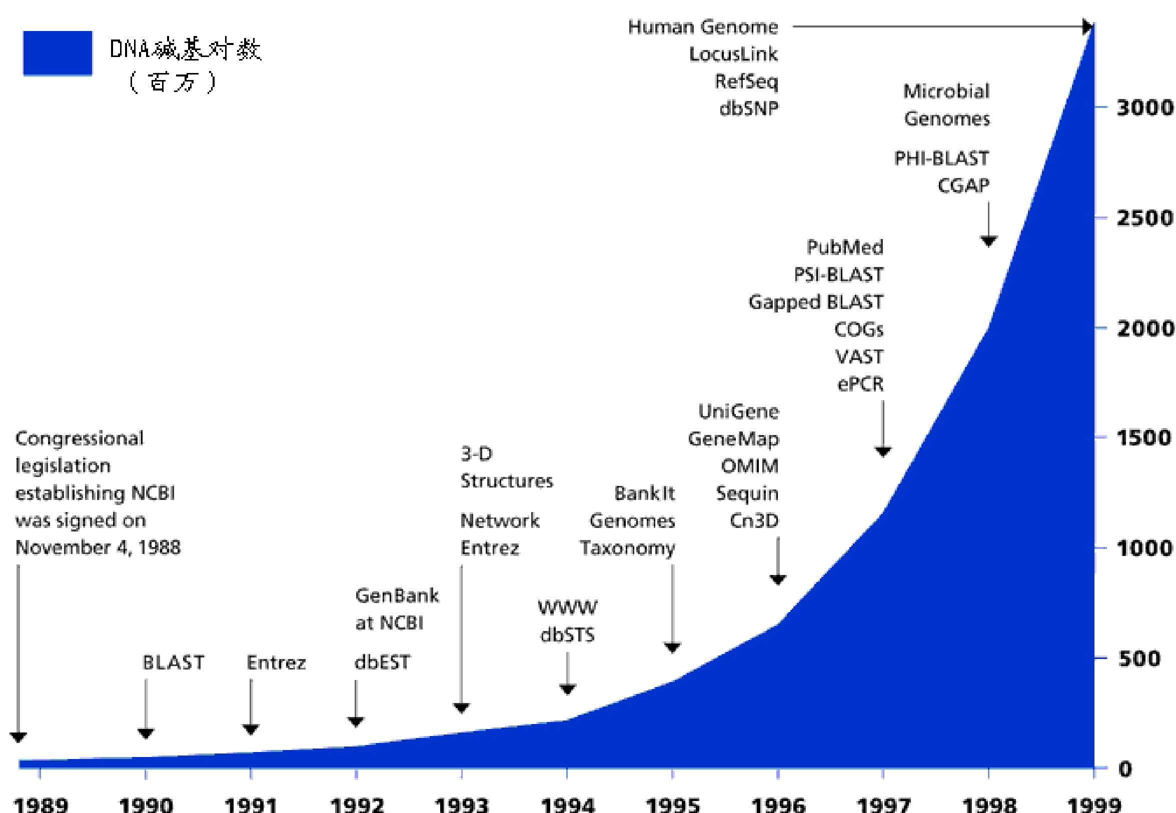


图 1.4　美华盛顿大学 Phil Green 教授。他所研制的自动测序组装系统 Phred-Phrap-Consed 被广泛应用于鸟枪法测序，其中包括人类基因组计划。



图 1.5　*Nature* 和 *Science* 2001 年 2 月 15 日和 16 日人类基因组专刊封面。*Science* 封面中的五位成年人分别为 Celera 公司人类基因组测序计划基因材料的提供者。

图 1.6 美国家生物技术信息中心(NCBI)十年(1989-1999)发展简史（NCBI，1999）



*图中涂黑部分表示 GenBank 数据库 DNA 碱基数据增长情况(单位：百万)；
**图中各年主要事件说明：
1989：NCBI 被国会批准于 1988 年 11 月成立；
1990：BLAST 搜索程序研制完成；
1991：Entrez 检索系统(光盘)建立；
1992：GenBank 划归 NCBI，NCBI 建立 EST(表达序列检签)数据库(dbEST)；
1993：Entrez 检索网络系统建立，同时 Entrez 中增加三维大分子结构数据内容；
1994：NCBI 网站建立，STS(序列标签位点)数据库(dbSTS)在 NCBI 建立；
1995：向 GenBank 发送 DNA 序列系统 BankIt 面市，随着人类基因组计划的开展和数据库数据的膨胀，NCBI 分别建立基因组数据库和分类浏览器；
1996：为了帮助从序列中发现基因，UniGene、GeneMap(人类基因转录图谱)、OMIM(Online Mendelian Inheritance in Man)、Cn3D 数据库建立，序列发送新系统 Sequin 面市；
1997：文献检索库 PubMed 上网,新的搜索程序 PSI-BLAST(Position-Specific Iterated BLAST)和 Gapped BLAST(允许空位)研制完成，载体搜索工具 VAST 和 PCR 分析软件 ePCR 面市，COG(Clusters of Orthologous Groups)开始用于蛋白质序列的直系同源分析；
1998：20 种微生物基因组数据被公开，PHI-BLAST(Pattern Hit Initialed BLAST)完成，癌症基因组结构计划(CGAP)开始实施；
1999：完成一系列用于人类基因组分析工具和资源：Locuslink、RefSeq 和 dbSNP。

Ouzounis 和 Valencia (2003) （见 Christos A. Ouzounis and Alfonso Valencia. Early bioinformatics: the birth of a discipline ------ a personal view. *Bioinformatics*. 2003, 19(17): 2176-2190）总结了截止到 10 年前（上世纪 90 年代初）生物信息学发展的重要研究成果，其中还列出了所谓“ TOP 20 PAPERS ”（表 1.3）。当然这只是他们的一家之言，难免有自己的偏好，仅供参考。例如著名的 Smith-Waterman 算法(1981)就没有被列入。

表 1.3 早期影响生物信息学发展的 20 篇经典文献。取自 Ouzounis and Valencia （2003）。

Publication	Comments
Zuckerkandl and Pauling, 1965b	First use of molecular sequences for evolutionary studies
Fitch and Margoliash, 1967	Use of molecular sequences to build trees
Needleman and Wunsch, 1970	First implementation of dynamic programming for protein sequence comparison
Lee and Richards, 1971	Calculation of accessibility on protein structures
Chou and Fasman, 1974	First secondary structure prediction method
Tanaka and Scheraga, 1975	Simulation of protein folding
Dayhoff, 1978	First collection of protein sequences
Hagler and Honig, 1978	One of the first explicit attempts to simulate protein folding
Doolittle, 1981	Seminal paper examining divergence and convergence in protein evolution
Felsenstein, 1981	One of the first statistical treatments of evolutionary tree construction
Richardson, 1981a	The most comprehensive description of protein structure to that date
Kabsch and Sander, 1984	Discovery with profound implications for model building by homology and structure prediction
Novotny *et al*., 1984	The inability of distinguishing correct from incorrect structures threw back structure prediction approaches for a long while
Chothia and Lesk, 1986	Examination of divergence between sequence and structure
Doolittle, 1986	Influential book on sequence analysis
Feng and Doolittle, 1987	The first approach for an efficient multiple sequence alignment procedure, later implemented in CLUSTAL
Lathrop *et al*., 1987	One of the first applications of Artificial Intelligence in protein structure analysis and prediction
Ponder and Richards, 1987	The very first threading approach, using sequence enumeration
Altschul *et al*., 1990	The implementation of a sequence matching algorithm based on Karlin's statistical work
Bowie *et al*., 1991	The first implementation of protein structure prediction using threading

第三节　基因组时代：生物信息学的应用与展望

蛋白质、DNA 和 RNA 序列的计算分析在上世纪 80 年代末已发生了根本性变化。高效实验新技术，特别是测序技术是这一变化的推动力，这些新技术使实验数据急剧增长。当基因组测序计划持续开展，研究重点已逐步从数据的积累转向数据的解释。用于序列分类、相似性搜索、DNA 序列编码区识别、分子结构与功能预测、进化过程的构建等方面的计算工具已成为研究工作的重要组成部分。这些工具有助于我们了解生命本质和进化过程，同时对新药和新疗法的发现具有重要意义。生物信息学已成为介于生物学和计算机科学学科前沿的重要学科，在许多方面影响着医学、生物技术和人类社会。现在作为一名分子生物学者，不具备一些基本的生物信息学技能已几乎难以胜

任。实验室的每一项技术，从简单的克隆、PCR 到基因表达分析都需要在计算机上进行数据的处理，这些工作均需要理解 DNA 和蛋白质分析工具的基本算法。



生物信息学家们面对的是堆集如山的 DNA 片段。这是在人类基因组序列 2001 年完成后出现的一幅漫画：如何真正破译人类自身的庞大的基因组？

我们处在一个基因组时代。许多新技术，如用于大规模测序工程的毛细电泳(capillary electrophoresis)，基因芯片制造的光刻技术(photolithography)和机器人技术(robotics technology)等应用于基因组研究，使我们能在以前不可能达到的尺度和角度上观察生物学现象：某一基因组的所有基因，某一个细胞中的所有转录产物，某一组织中的所有代谢过程。这些新技术的一个共同特点是产生大量的数据。例如GenBank数据库已拥有了超过 10^{10}个DNA序列数据，并以每年翻一翻的速度增长。那些分析基因表达模式、蛋白质结构、蛋白质间互作等的新技术又会产生更多的数据。如何管理这些数据、解读它们并使各领域的生物学家们能容易地使用它们是生物信息学面临的巨大挑战。

生物信息学面临着越来越多的困难，许多困难是在我们面对大规模科技工程时，所有生物学家都将碰到的问题。对初学者而言，很少有人能在计算机科学和生物学研究两方面同时拥有扎实的背景。这一问题将使那些可以培养下一代生物信息学者的人才匮乏。同时，对对方研究问题的无知可能导致误解。例如，编写用于拼接 EST 重叠群的程序对于生物学者来说是非常重要的，但对于计算机科学家来说，这没有任何新意。同样，证明在一定条件下不可能构建一个整体最佳系统树(phylogenetic tree)可能是计算机科学的一个重要命题，但对于生物学家来说并无什么实践意义。如何找到共同感兴趣的问题是生物信息学的重要目标。所谓“真正”的生物学研究已越来越多地在计算机前完成，同时，越来越多的计算机科学的课题将来自生物学问题。

一个生物信息学研究者需要怎样的基本条件呢？Gibas and Jambeck 在他们的《Developing Bioinformatics Computer Skills》(C. Gibas and P. Jambeck, O’REILLY, 2001)书中大致给出了如下标准：

•You should have a fairly deep background in some aspect of molecular biology. …but without a core of knowledge of molecular biology you will, as one person told us, “run into brick walls too often.”
•You must absolutely understand the central dogma of molecular biology.
•You should have substantial experience with at least one or two major molecular biology software packages, either for sequence analysis or molecular modeling.
•You should be comfortable working in a command-line computing environment.
•You should have experience with programming in a computer language such as C/C++, as well as in a scripting language such as Perl or Python.

生物信息学作为一个组合学科，需要有多方面的数据资源，这无疑又增加了生物信息学面临的困难。没有这些数据资源和以新方式组合这些数据的能力，生物信息学学科领域范围将受到极大限制。例如，基因相似性搜索程序 BLAST，它的广泛应用除了得益于它的算法外，还得益于那些公共数据库，如 GenBank、EMBL 和 DDBJ。没有这些数据库供查询，BLAST 将作用有限。

生物信息学研究的一个核心问题是数据库的开发：如何整合和最有效地查询来自诸如基因组 DNA 序列、mRNA 表达的空间和时间模式(spatial and temporal pattern)、蛋白质结构、免疫反应、文献记录等数据。其次是从诸如组装完成的核酸或蛋白质序列中识别模式的算法、用于相似性比较或系统发育构建的序列列线(alignment)、线性序列或高维结构的模序(motif)识别和基因表达的共有模式等等。

如上所述，数据的共享性和应用性非常重要，这引起人们对数据释放(公开)政策的关注：初级数据(primary data)的组成、谁应拥有这些数据、应什么时候和如何公开、对数据的进一步使用可否设置限制等。目前已经隐现的两方面问题可能阻碍生物信息学研究的进展，即(1)数据公开前的使用问题和(2)对已公开数据的保存限制。认识到数据尽早释放对许多研究具有重要意义，人类基因组计划(Human Genome Project，HGP)采用了一种数据正式公布前即上网释放的政策，许多其它基因组计划目前也采用了相同的做法。由于生物信息学强烈依赖于各种来源的数据资源，所以希望一些基因组水平的研究计划(如表达分析和蛋白质组学研究)也能采取相同的政策。但是，这种利他主义的数据释放政策需要一些保护，如对产生初级数据的人应能使之得到应有的认可。有人最近提出用类似于“ 私人通信 ”的方式来处理这些尚未正式公布的数据，这样可以在一定程度上保护这些数据的知识产权。生物信息学研究面对的第二个问题并不是对数据使用的限制而是对下游研究的限制，如将一些数据并入新的或已有的数据库中。这一问题对于生物信息学研究更为关键，因为这不仅涉及何时可以进行生物信息学分析并可进行何种分析。塞莱拉(Celora)公司最近公布的人类基因组初步分析结果便集中引发了这一问题。该公司测得的原始数据(即初级数据)仅由这家私人公司释放，并对这些数据的进一步存储和加工设定了限制。不妨想象一下，基因组学研究处于这样一种境地，公共数据库(GenBank/EMBL/DDBJ)没有相应数据，由于所有权的限制使数据拼接无法进行。5 年前，百慕大协定(Bermuda Conventions)为基因组序列的释放建立了一个很好的标准；鉴于数据释放和使用政策对生命科学研究的深远影响，我们有必要认真考虑为接下来的 5 年制定些什么标准。在后基因组时代(postgenomic era)，人们期待在对生物发育机理、代谢过程和疾病认识方面有所突破。可以肯定地预言，生物信息学研究将对我们的一些认识产生根本性改变，如基因表达调控、蛋白质结构预测、比较进化学和药物开发等领域。只有在数据共享的情况下，基因组水平的研究才有可能进行。捆住手脚，要在数据的海洋中畅游是很困难的。

在中国，生物信息学随着人类和水稻等基因组研究的展开已显露出蓬勃发展的势头。许多大学和科研院所已经投入大量人力开设生物信息学专业、建立生物信息学研

究所（中心）并从事这方面的研究工作，例如北京大学生物信息中心(http://www.ipc.pku.edu.cn/)、中国科学院上海生命科学院生物信息中心(www.biosino.org.cn)、清华大学、天津大学、内蒙古大学、复旦大学以及浙江大学生物信息学研究所（http://ibi.zju.edu.cn）等等。生物信息学作为基因研究的有力武器，被广泛用于新基因的发现，以达到将有用新基因抢先注册专利的目的。在这场抢基因的国际竞争中，如何结合我国科研、开发状况，重点投入以求得局部优势和商业回报，是中国科学家和相关部门必须面对的新课题。

第二章　分子数据库

生物信息学涉及的数据库可大致分为二种：初级数据库和二级数据库。初级数据库贮存原始的生物数据，如 DNA 序列，由晶体衍射(Crystallography)获得的蛋白质结构等。二级数据是在初级数据库的基础上经加工和增加相关信息，使它们更便于特定专业人员的使用，如真核生物启动子序列库 EPD 和蛋白质一般结构或功能模体(motif)数据库 PROSITE。

一个数据库记录(entry)一般由两部分组成：原始序列数据和描述这些数据生物学信息的注释(annotation)。注释中包含的信息与相应的序列数据同样重要和有应用价值，这一点值得注意。在基因组规模上的测序过程便产生了注释问题。对于那些从自动测序仪中出来的序列，我们往往只知道它们来自何种细胞类型，而其它方面却知之甚少。如果你在确定一段未知蛋白质序列的功能，发现一个与之匹配的序列，但该序列却没有任何有关功能的信息时，你的研究工作便很难为继了。

不同的数据库的注释质量差异很大，因为一个数据库往往要在数据的完整性和注释工作量之间寻找一个平衡点。一些数据库提供的序列数据很广，但这必定会影响序列的注释；相反，一些数据库数据面较窄，但它提供了非常全面的注释。数据库记录的注释工作是一个动态过程，新的发现不断被补充进去，所以，本书中用到的一些注释信息可能很快便被更新了。在所有的生物信息数据库中总会有一小部分的记录(包括原始序列数据和注释)是不正确的，这是一个无法避免的事实。

第一节　　初级数据库

一．DNA 数据库

DNA 序列构成了初级数据库的主体部分。目前国际上有 3 个主要的 DNA 序列公共数据库(表 2.1)：欧洲分子生物学实验室(European Molecular Biology Laboratory, EMBL)(位于英国剑桥)，GenBank[美国国家生物技术信息中心(National Center for Biotechnology Information, NCBI)，该中心隶属于美国家医学图书馆，位于美国国家卫生研究院(NIH)内]和日本 DNA 数据库(DNA Databank of Japan,DDBJ)。这 3 个大型数据库于 1988 年达成协议，组成合作联合体。它们每天交换信息，并对数据库 DNA 序列记录的统一标准达成一致。每个机构负责收集来自不同地理分布的数据(EMBL 负责欧洲,GenBank 负责美洲,DDBJ 负责亚洲等)，然后来自各地的所有信息汇总在一起，3 个数据库共同享有并向世界开放，故这 3 个数据库又被称为公共序列数据库（Public Sequence Database)。所以从理论上说，这 3 个数据库所拥有的 DNA 序列数据是完全相同的。你可以从中选择一个你喜欢的数据库；但是如果你的研究需要实时(24 小时以内)的，则要注意这些数据库间的记录是会有差异的。

表 2.1 三个主要 DNA 序列数据库网址

数据库（Database）	网址（Address）
EMBL	www.ebi.ac.uk/ebi_docs/embl_db/ebi/topembl.html
GenBank	www.ncbi.nlm.nih.gov/Genbank/GenbankSearch.html
DDBJ	www.ddbj.nig.ac.jp

DNA 序列数据库的增长是飞速的。EMBL 核酸数据近 20 年的增长情况(表 2.2)充分说明了这一点。从历史看，每 22 个月，数据库的数据规模将翻一翻，而且随着表达序列标签(EST)数据的迅猛增长，这一速率已快速递增。EMBL 数据翻一翻的周期最近已缩短到 9 个月左右。数据库的膨胀对于数据库的搜索非常有好处。也许你这个月还找不到一个匹配序列，但可能在下一次更新的数据中寻获。所以，当进行生物信息学分析时，分析结果中务必要注明你当时所使用序列数据库的数据状况。2004 年 12 月 EMBL（Release81)的 DNA 碱基对数已接近 800 亿，序列数超过 4 亿条。为了有效地管理如此庞大的数据，数据库数据根据物种(species)分为几类，每个记录都被严格地归入某一类中。每一类用了 3 个字母代码表示(表 2.3)，例如 EMBL 数据库最近的分类为人、真菌等等，同时每一类的数据文件往往又分成一定的亚类，例如 EST 类数据文件（表 2.4)。这些分类有助于我们便捷地进入数据库的相关部分。这些分类并非一成不变，随着时间的推移可能进行一定的修正，这主要是数据库规模快速扩大的需要。如新加入的高通量测序数据(HTG)等。EMBL 和 GenBank 等数据库的使用手册均可在相应的网址上找到，这些手册提供了详尽的数据库组成、分类等细节，不妨到那些网站(见表 2.1)上看看。

表 2.2 EMBL 数据库 DNA 序列数据库增长情况

数据库报告 (Release)	释放日期 (Month)	记录数 (Entries)	核苷酸数 (Nucleotides)
Release 1	1982 年 6 月	568	585433
Release 7	1985 年 12 月	5789	5622638
Release 25	1990 年 11 月	41580	52900354
Release 29	1991 年 12 月	57655	75400487
Release 33	1992 年 12 月	89100	111413979
Release 37	1993 年 12 月	146576	158171400
Release 41	1994 年 12 月	230950	226259607
Release 45	1995 年 12 月	622566	427620278
Release 49	1996 年 12 月	1047263	696183789
Release 53	1997 年 12 月	1917868	1281391651
Release 57	1998 年 12 月	3046471	2164718256
Release 61	1999 年 12 月	5303436	4508169737
Release 65	2000 年 12 月	9549328	10710321435
Release 69	2001 年 12 月	14366182	15383451165
Release 73	2002 年 12 月	20857746	27903283528
Release 77	2003 年 12 月	30351263	36042464651
Release 81	2004 年 12 月	46105397	79271300840

表 2.3 EMBL 数据库 2004 年 12 月数据状况 (Release81)

类	别 (Division)	代码 (Code)	记录数 (Entry)	核苷酸数 (Nucleotide)
表达序列标签	ESTs	EST	24481418	12837493911
真菌	Fungi	FUN	110405	221397562
基因组检测序列	Genome Survey Sequences	GSS	10726912	6608825736
高通量基因组	High Throughput Genome	HTG	68564	11613533555
人	Human	HUM	292205	4126190851
无脊椎动物	Invertebrates	INY	175545	677544114
其它哺乳动物	Other Mammals	MAM	70355	341455910
细胞器	Organelles	ORG	314215	270405172
专利	Patents		2276431	1332968224
噬菌体	Bacteriophage	PHG	2625	12989224
植物	Plants	PLN	287510	1084488061
原核生物	Prokaryotes	PRO	282227	993811176
啮齿类动物	Rodents	ROD	31538	110601526
序列标签位点	STSs	STS	380660	168545968
合成	Synthetic	SYN	14240	22721647
未分类	Unclassified	UNC	2869	2823924
病毒	Viruses	VRL	262346	241496438
其它脊椎动物	Other Vertebrates	VRT	113601	879447919
总和	Total		39893666	41546740918

表 2.4 EMBL 数据库 EST 类数据文件分类情况(Release 81)

亚类数据文件名 (Subdivision)		说明 (Comments)
est_fun.dat	est_fun05.dat	真菌 EST
est_hum.dat	est_hum57.dat	人 EST
est_inv.dat	est_inv31.dat	无脊椎动物 EST
est_mam.dat	est_mam11.dat	哺乳动物 EST
est_pln.dat	est_pln55.dat	植物 EST
est_pro.dat	est_pro01.dat	原核生物 EST
est_rod.dat	est_rod07.dat	啮齿类动物 EST
Est_vrt.dat	est_vrt27.dat	脊椎动物 EST

二．基因组数据库

第二个主要的初级数据源来自各种基因组计划。一些基因组计划已经完成，如真核生物酵母(*Saccharomyces cerevisiae*)，原肠生物(*Methanococcus janeschii*) 和 3 个原核生物流感嗜血杆菌(*Haemophilus influenzae*)、(*Mycoplasma genitaliam*)和大肠杆菌(*Escherichia coli*)等。这些计划的大部分信息在 EMBL 中均可找到。很多基因组计划正在进行中，表 2.5 列出了一些基因组计划的网址。

表 2.5 部分生物基因组计划网址

生物种类	Organism	网址(Address)
曲霉菌	Aspergillus	http://www.ncbi.nlm.nih.gov/genome/guide/aspergillus
蜜蜂	Bee	http://www.ncbi.nlm.nih.gov/genome/guide/bee
猫	Cat	http://www.ncbi.nlm.nih.gov/genome/guide/cat
青蛙	Frog	http://www.ncbi.nlm.nih.gov/genome/guide/frog
老鼠	Mouse	http://www.ncbi.nlm.nih.gov/genome/guide/mouse
小鼠	Rat	http://www.ncbi.nlm.nih.gov/genome/guide/rat/index.html
狗	Dog	http://www.ncbi.nlm.nih.gov/genome/guide/dog
牛	Cow	http://www.ncbi.nlm.nih.gov/genome/guide/cow
猪	Pig	http://www.ncbi.nlm.nih.gov/genome/guide/pig
羊	Sheep	http://www.ncbi.nlm.nih.gov/genome/guide/sheep
鸡	Chicken	http://www.ncbi.nlm.nih.gov/genome/guide/chicken
斑马鱼	Zebra fish	http://www.ncbi.nlm.nih.gov/genome/guide/zebrafish/index.html
海胆	Sea urchin	http://www.ncbi.nlm.nih.gov/genome/guide/sea_urchin
线虫	Caenorhabditis elegans	http://www.ncbi.nlm.nih.gov/genome/guide/nematode
	Dictyostelium discoideum	http://www.ncbi.nlm.nih.gov/genome/guide/dicty
果蝇	Drosophila	http://www.ncbi.nlm.nih.gov/genome/guide/fly
蚊子	Mosquito	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?chr=agambiae.inf
黑猩猩	Chimp	http://www.ncbi.nlm.nih.gov/genome/guide/chimp
人	Human	http://www.ncbi.nlm.nih.gov/genome/guide/human
拟南芥	Arabidopsis	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=3702
棉花	Cotton	http://algodon.tamu.edu
玉米	Maize	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4577
水稻	Rice	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4530
小麦	Wheat	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4565
大麦	Barley	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4513
大豆	Soybean	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=3847
西红柿	Tomato	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4081
高粱	Sorghum	http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=4557

生物基因组差异极大，这种差异不仅表现在基因组大小(表 2.6)而且在于单链或双链 DNA、RNA 的遗传信息存储特性。另外，一些基因组是线性的(如哺乳动物)，而另一些则上封闭的环状(如绝大多数细菌)。

人类最早(1977)获得的生物基因组全序列是噬菌体（53kb)，1987 年自动测序仪问世，随后第一个病毒基因组序列（1990）在自动测序仪上完成；后来是第一个细菌基因组(1995)被完全测序，紧接着是酵母(1996)、多细胞线虫(1998)和果蝇(1999)基因组，最后是人类自身(2000)的遗传密码被解开。最早完成的噬菌体、病毒和细胞器的基因组数据在 80 年代早期就存入了 EMBL 数据库。从那以后，随着测序技术的革命性改进，大量的基因组数据被存入该数据库，涉及的物种种类不断增多，如最近又增加了（Chimp）和（Fruit Fly)。表 2.6 列举了一些生物基因组测序的进展状况。欧洲生物信息学研究所(European Bioinformatics Institute，EBI)的基因组网站提供了已完成的基因组序列数据，可以自由访问(www.ebi.ac.uk/genomes/)（图 2.1)。一些网站提供世界范围内基因组测序进展的最新情况，如 Genome MOT（图 2.2)，通过它们可以了解基因组测序的发展动态。

2000 年 6 月 26 日，人类基因组草图被宣告完成。原始的序列数据可在 EMBL 等数据库的 HTG 和 HUM 部分找到。人类基因组的大小为 3.2 兆亿碱基对(Gigabases)，而由于冗余原因，实际获得的碱基数超过了这一数字。总的估计，冗余的碱基比例为 30-40%。

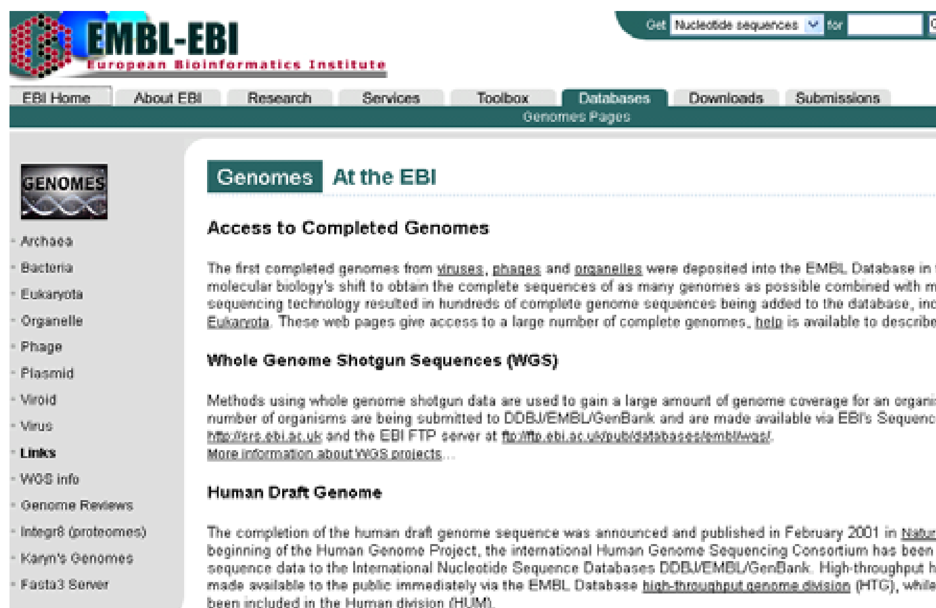


图 2.1 欧洲生物信息学研究所(EBI)的基因组网站主页。该网站提供了已完成的各类生物（真核生物、细菌、病毒等，见图中左列）基因组情况。

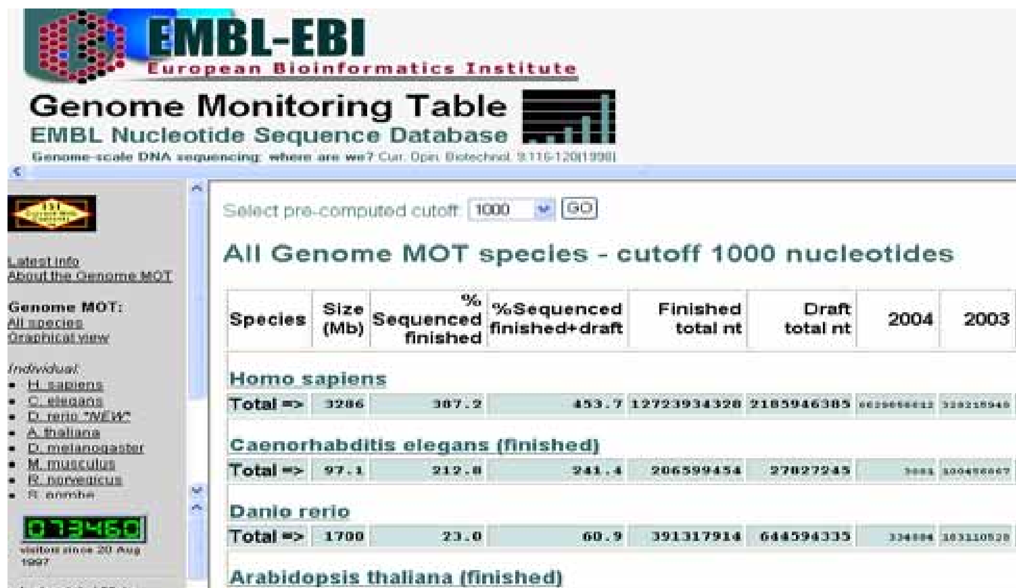


图 2.2 基因组测序进展状况服务器 Genome MOT 主页。该网站每日报告基因组测序进展情况，图中为 2004 年 2 月 29 日的进展报告，具体数据可参见表 2.6。

表 2.6 部分基因组测序情况。根据 Genome MOT，截止 2004/2/29。

物种 Species	基因组大小 Size (Mb)	完成比率 %Sequenced finished	完成比率（包括草图）% Sequenced finished+draft	完成核苷酸总数 Finished total nt	完成(草图)核苷酸总数 Draft total nt
人类 *Homo sapiens*	3286	387.2	453.7	12723934328	2185946385
线虫 *Caenorhabditis elegans* (finished)	97.1	212.8	241.4	206599454	27827245
拟南芥 *Arabidopsis thaliana* (finished)	118	205.0	207.6	241889120	3060859
果蝇 *Drosophila melanogaster* (finished)	135.6	324.0	432.3	439303114	146899025
Danio rerio	1700	23.0	60.9	391317914	644594335
老鼠 *Mus musculus*	3059	134.0	197.4	4099660644	1938348462
小鼠 *Rattus norvegicus*	3000	1.8	170.6	52581865	5066240322
Schizosaccharomyces pombe (finished)	13.8	205.7	205.7	28386200	0
酵母 *Saccharomyces cerevisiae* (finished)	12.1	274.9	274.9.7	33258957	0

三．蛋白质序列数据库

SWISS-PROT 和 PIR 是国际上二个主要的蛋白质序列数据库，目前这二个数据库在 EMBL 和 GenBank 数据库上均建立了镜像（mirror）站点。SWISS-PROT 数据库包括了从 EMBL 翻译而来的蛋白质序列，这些序列经过检验和注释。该数据库主要由日内瓦大学医学生物化学系和欧洲生物信息学研究所(EBI)合作维护。SWISS-PROT 的序列数量呈直线增长。SWISS-PROT 的数据存在一个滞后问题，即把 EMBL 的 DNA 序列准确地翻译成蛋白质序列并进行注释需要时间。一大批含有开放阅读框(ORF) 的 DNA 序列尚未列入 SWISS-PROT。为了解决这一问题，TREMBL(Translated EMBL)被建立了起来。TREMBL 也是一个蛋白质数据库，它包括了所有 EMBL 库中的蛋白质编码区序列，提供了一个非常全面的蛋白质序列数据源，但这势必导致其注释质量的下降。PIR 数据库的数据由美国家生物技术信息中心(NCBI)翻译自 GenBank 的 DNA 序列。PIR 根据注释程度(质量)分为 4 个等级(表 2.7)。

表 2.7 PIR 数据库的分类情况 (Release 80)

分类名称 (Name)	说明 (Comment)	记录数 (Number of entries)
PIR1	分类并注释 (Classified and annotated)	20685
PIR2	注释（Annotated ）	262300
PIR3	未核实（Unverified ）	24
PIR4	未翻译 (Unencoded or untranslated)	407

表 2.8 列出了以上主要蛋白质序列数据库的网址，有关详情可到这些网站上获得。

表 2.8 主要蛋白质序列数据库网址

数据库 (Database)	网址 (Address)
SWISS-PROT	http://www.ebi.ac.uk/swissprot/
TREMBL	http://www.ebi.ac.uk/trembl/
PIR	http://pir.georgetown.edu/

4．蛋白质结构数据库

实验获得的三维蛋白质结构均贮存在蛋白质数据库 PDB 中。PDB 是国际上主要的蛋白质结构数据库，虽然它没有蛋白质序列数据库那么庞大，但其增长速度很快。PDB 贮存有由 X 射线和核磁共振(NMR)确定的结构数据。NRL-3D 数据库提供了贮存在 PDB 库中蛋白质的序列，它可以进行与已知结构的蛋白质序列的比较。对来自 PDB 中每个已知三维结构的蛋白质序列进行多序列同源性比较(multiple sequence alignment)的结果，被贮存在 HSSP(homology-derived structures of proteins)数据库中。被列为同源的蛋白质序列很有可能具有相同的三维结构，HSSP 因此根据同源性给出了 SWISS-PROT 数据库中所有蛋白质序列最有可能的三维结构。要想了解对已知结构蛋白质进行等级分类的情况可利用

SCOP(Structural classification of proteins)数据库，在该库中可以比较某一蛋白质与已知结构蛋白的结构相似性。CATH 是与 SCOP 类似的一个数据库。

表 2.9 主要蛋白质结构数据库网址

数 据 库 (Database)	网 址 (Address)
PDB	http://www.rcsb.org/pdb
NRL-3D	http://pir.georgetown.edu/pirwww/search/textnrl3d.html
HSSP	http://www.sander.embl-heidelberg.de/hssp
SCOP	http://scop.mrc-lmb.cam.ac.uk/scop
CATH	http://www.biochem.ucl.ac.uk/bsm/cath

第二节 初级序列数据的注释

到目前为此，尚没有一个统一的序列注释格式，各数据库间均存在差异。但总的来说，各数据库所提供的注释内容还是相同的。现在比较使人不放心的是针对一个相同基因的 DNA 和蛋白质序列注释之间的差异。以下给出了一个 EMBL 数据库记录的注释例子(图 2.2)。表 2.10 对注释中的代码及内容进行了说明。

```
ID   LISOD      standard; DNA; PRO; 756 BP.
XX
AC   X64011; S78972;
XX
SV   X64011.1
XX
DT   28-APR-1992 (Rel. 31, Created)
DT   30-JUN-1993 (Rel. 36, Last updated, Version 6)
XX
DE   L.ivanovii sod gene for superoxide dismutase
XX
KW   sod gene; superoxide dismutase.
XX
OS   Listeria ivanovii
OC   Bacteria; Firmicutes; Bacillus/Clostridium group;
OC   Bacillus/Staphylococcus group; Listeria.
XX
RN   [1]
RX   MEDLINE; 92140371.
RA   Haas A., Goebel W.;
RT   "Cloning of a superoxide dismutase gene from Listeria ivanovii by
RT   functional complementation in Escherichia coli and characterization of the
RT   gene product.";
RL   Mol. Gen. Genet. 231:313-322(1992).
XX
RN   [2]
RP   1-756
RA   Kreft J.;
RT   ;
RL   Submitted (21-APR-1992) to the EMBL/GenBank/DDBJ databases.
RL   J. Kreft, Institut f. Mikrobiologie, Universitaet Wuerzburg, Biozentrum Am
RL   Hubland, 8700 Wuerzburg, FRG
XX
DR   SWISS-PROT; P28763; SODM_LISIV.
XX
FH   Key             Location/Qualifiers
FH

FT   source          1..756
FT                   /db_xref="taxon:1638"
FT                   /organism="Listeria ivanovii"
FT                   /strain="ATCC 19119"
FT   RBS             95..100
FT                   /gene="sod"
FT   terminator      723..746
FT                   /gene="sod"
FT   CDS             109..717
FT                   /db_xref="SWISS-PROT:P28763"
FT                   /transl_table=11
FT                   /gene="sod"
FT                   /EC_number="1.15.1.1"
FT                   /product="superoxide dismutase"
FT                   /protein_id="CAA45406.1"
FT                   /translation="MTYELPKLPYTYDALEPNFDKETMEIHYTKHHNIYVTKLNEAVSG
FT                   HAELASKPGEELVANLDSVPEEIRGAVRNHGGGHANHTLFWSSLSPNGGGAPTGNLKAA
FT                   IESEFGTFDEFKEKFNAAAAARFGSGWAWLVVNNGKLEIVSTANQDSPLSEGKTPVLGL
FT                   DVWEHAYYLKFQNRRPEYIDTFWNVINWDERNKRFDAAK"
XX
SQ   Sequence 756 BP; 247 A; 136 C; 151 G; 222 T; 0 other;
     cgttatttaa ggtgttacat agttctatgg aaatagggtc tatacctttc gccttacaat        60
     gtaatttctt ttcacataaa taataaacaa tccgaggagg aatttttaat gacttacgaa       120
     ttaccaaaat taccttatac ttatgatgct ttggagccga attttgataa agaaacaatg       180
     gaaattcact atacaaagca ccacaatatt tatgtaacaa aactaaatga agcagtctca       240
     ggacacgcag aacttgcaag taaacctggg gaagaattag ttgctaatct agatagcgtt       300
     cctgaagaaa ttcgtggcgc agtacgtaac cacggtggtg gacatgctaa ccatacttta       360
     ttctggtcta gtcttagccc aaatggtggt ggtgctccaa ctggtaactt aaaagcagca       420
     atcgaaagcg aattcggcac atttgatgaa ttcaaagaaa aattcaatgc ggcagctgcg       480
     gctcgttttg gttcaggatg ggcatggcta gtagtgaaca atggtaaact agaaattgtt       540
     tccactgcta accaagattc tccacttagc gaaggtaaaa ctccagttct tggcttagat       600
     gtttgggaac atgcttatta tcttaaattc caaaaccgtc gtcctgaata cattgacaca       660
     ttttggaatg taattaactg ggatgaacga aataaacgct ttgacgcagc aaaataatta       720
     tcgaaaggct cacttaggtg ggtcttttta tttcta                                  756
```

图 2.3 EMBL 数据库记录(记录是 X64011)注释例举。有关说明见表 2.10

表 2.10 EMBL 数据库记录注释代码和内容说明

代码 (Code)	全 称 (Full meaning)		说 明 (Comments)
ID	identifier	(身份号)	该行的第一项内容是该数据库记录的名称，该名称是唯一的，是由 EMBL 数据库给定的。其它内容注明了该记录的一些状况(如是否已经被核实－本例中为已核实，即 standard；记录的碱基数等)
AC	accession number	(记录号)	每个记录号均是唯一的，并从不更改，是由 GenBank 给定的。如果两个记录被合并成一个记录，原始上着 2 个记录号均会被注明
DT	data	(日期)	2 个日期被注出，一个是该数据第一次被记录时间，另一个是最后一次的时间。
DE	description	(描述)	对该基因的文字描述
KW	keywords	(关键词)	描述该基因的关键词
OS	organism(species)	(物种)	物种名称
OC	organism(classification)	(分类)	物种的一个简单分类，该分类并不一定准确，应谨慎从事
OG	Organelle	(细胞器)	该基因是否在某一个特殊的细胞器中
RN	reference number	(文献编号)	
RC	reference comment	(文献说明)	
RP	reference positions	(文献大小)	与该记录研究相关的文献信息
RX	cross-reference	(相关文献)	
RA	reference authors	(文献作者)	
RT	reference title	(文献题目)	
RL	reference location	(文献出处)	
DR	database cross-reference	(相关文献数据库)	见文中说明
FH	feature header	(主表头)	该记录主要内容列表表头
FT	feature table data	(主表数据)	见文中说明
CC	comments	(说明)	对记录的文字说明
XX	spacer line	(空白行)	
SQ	sequence header	(序列头)	有关该序列大小和组成的信息
blank	sequence data	(空白)	
//	termination line	(终止行)	一个记录的终止符号

相关文献数据库 (database cross-reference，DR)需要做进一步的说明。许多二级数据库内容来自初始数据库，例如 OMIM(Online Mendelian Inheritance in Man)数据库是有关人类遗传疾病的数据，如果 OMIM 中的一个记

录与 EMBL 中一个已知序列的基因有关，则该基因将与该记录建立联系，则 EMBL 库中该序列的 DR 栏中将包括 OMIM 和 OMIM 中相关记录的名称。上述例子(图 2.3)的 DR 栏中有该 DNA 序列翻译成蛋白质序列的 SWISS-PROT 记录号等。从此可见，DR 栏内容非常重要，它有助于了解与该原始 DNA 序列相关信息的状况和存贮站点。与 DR 栏可能有关的一些数据库包括 SWISS-PROT、 EMBL、OMIM、PROSITE(保守蛋白质模序数据库，见下文)、HSSP、PDB、MEDLNE(与 RL 栏相关的文献摘要数据库)、PIR 等。注释中另一个需要说明的重要内容是主表数据(feature table data, FT)栏。主表试图将尽可能多的序列信息囊括其中，并以计算机可以阅读的格式编排。3 个主要 DNA 数据库(EMBL、GenBank 和 DDBJ)已经对该表的表述格式达成了一致。具体表述格式内容说明可在 www.ebi.ac.uk/ebi_docs/embl_db/ft/feature_table.html 找到。

大量的 DNA 序列记录包含有一个以上的开放读框(ORF)。主表中的 PID 编号被用于唯一地指定每一个 ORF。这一编号是一个非常重要的注释信息，因为它可以使许多不同的 SWISS-PROT 记录与一个相同的 EMBL 序列相链接，可以精确地知道 EMBL 序列中的 ORF 所对应的 SWISS-PROT 蛋白质记录。

第三节 数据库信息检索系统

许多系统可以为使用者提供简便的序列库信息查寻服务，其中最著名和操作性最强的 2 个系统是 Entrez(由美国建立)和 SRS(Sequence retrieval System)(由 EMBL Theore Etzold 建立)。

SRS 检索系统在欧洲的许多网站被广泛使用。SRS 是一个具有弹性的系统，可应用于大量不同的数据库。这意味着使用 SRS 的数据库在各个站点可能略有差异，而这种差异是由数据库管理者所决定的。例如，OWL 数据库是一个非冗余蛋白质序列库，它的数据来源主要是从其它主要蛋白质数据库中收集而来的，在 SEQNET 服务器(www.seqnet.dl.ac.uk/srs/srsc)可通过 SRS 搜索而进入 OWL，但在 EBI 网站通过 SRS 则不能进入 OWL。

序列一般可通过记录号(如来自 1 篇发表的论文)或是该序列注释中的一些信息进行检索。SRS 的优势是可以使你通过普通的终端去检索大范围的数据库，并通过 DR 栏链接到在其它数据库。

SRS 的使用非常直观。图 2.4 所示是 EBI 网站 SRS 的主网。如果想检索序列数据可选择第一按钮"Search sequence libraries"。这时出现一个新图面(图 2.5)，然后选定你想搜索的数据库并在文字框(text)中键入正确的检索词。检索可建立逻辑关系(and,or,not)进行。按下"DO-QVERY"按钮便开始检索。检索结果的输出格式等也可设定，选定的记录内容可通过网络浏览器上的保存功能存入你的计算机中。

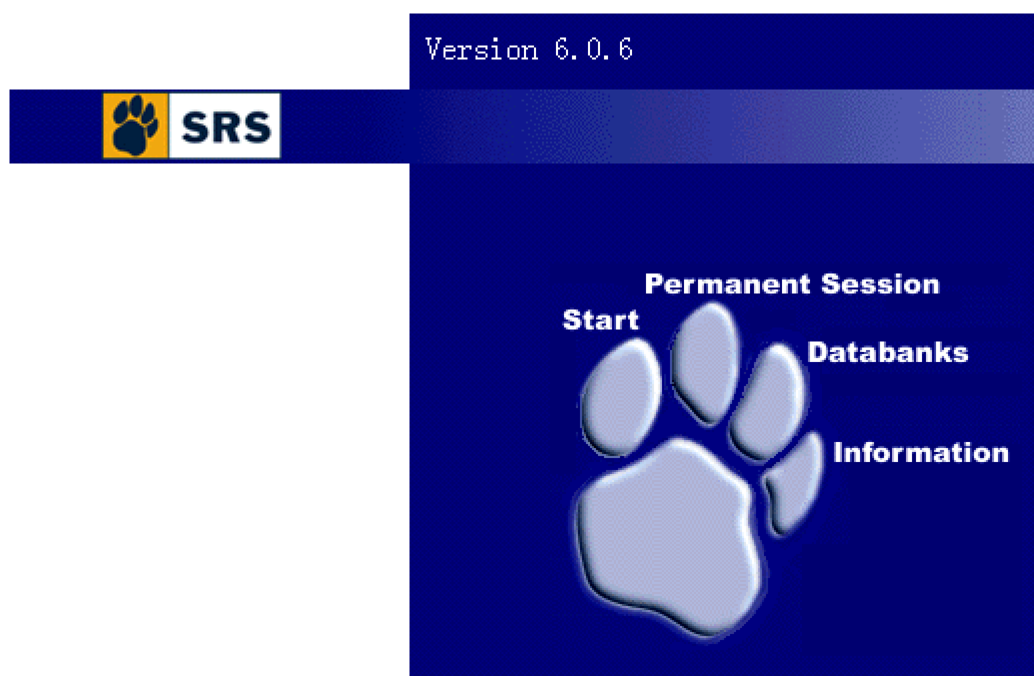


图 2.4 数据库搜索系统 SRS 主页

TOP PAGE QUERY RESULTS SESSIONS VIEWS DATABANKS HELP

Reset search EMBL Info about field AllText

Submit Query

separate multiple values by & (and), | (or), ! (and not)

append wildcards to words

AllText
AllText
AllText
AllText

combine searches with AND

retrieve entries of type Entry

Number of entries to display per page 30

Use predefined view * Names only *

Create your own view

ID
Division
AccNumber
SeqVersion
NID
Molecule
Description

Display in List Table

Extended query form

sequence format embl

图 2.5 数据库搜索系统 SRS 进行序列搜索设置的页面

第四节 数据库的冗余与偏差[1]

在进行 DNA 和蛋白质序列分析时碰到的一个棘手问题是数据库的冗余(redundancy)。DNA 和蛋白质数据库中的很多记录是属于同一基因和蛋白质家族，或在不同生物体上发现的同源基因。不同的研究机构可能向数据库发送了相同的序列数据，如果没有被检查出来，则这些记录或多或少地紧密相关。当然，这些记录如果的确非常相近，可以被认定为它们是相同序列，但一些显著的差异可能是由于基因组多样性的结果。

冗余数据至少可能导致以下 3 个潜在的错误：一是如果一组 DNA 或氨基酸序列包含了大量非常相关序列族，则相应的统计分析将偏向这些族，在分析结果中，这些族的特性被夸大；二是序列间不同部分的显著相关可能是在数据样本抽样时是有偏的和不正确的；最后是如果这些数据是被用于预测，则这些序列将使预测方法—如人工智能方法－发生偏离。

基于以上原因，有必须避免在数据库中存在太过于相似的序列，很多数据库也是这样做了，努力使他们的数据库为非冗余(non-redundant,nr)。但是，生物数据非常复杂，它远非“冗余”二字可以准备描述，例如，同一位点上的 2 个等位基因是不是冗余的？同一生物体内的 2 个同功酶是否冗余？因此，过于苛刻地去除“太过于相似的序列”可能导致一些有价值的信息被删除，应在数据规模和非冗余之间找到一个合理的平衡点。“太过于相似”的准确界定应主要依据所要研究的问题。实际研究中，试验数据往往“随机”地从数据库中抽取而不考虑减少冗余问题；即使考虑到冗余问题，也或多或少地存在随意性，即随意地进行一些同源性分析，确定一些蛋白质或 DNA 簇，然后从各簇中选取一个数据样本来组合所谓的“代表性”数据样本。

序列数据的偏差或错误(artifacts) 主要来自实验过程，这与其它科学数据的情况相同。这些错误主要来自以下几个方面：

(1) 载体序列污染：在测序列等实验过程中，载体序列可能造成污染，致使序列记录数据中包含了载体序列；
(2) 异源(heterologous)序列污染：有研究表明一些人类 cDNA 测序结果在实验过程中被酵母和细菌序列污染；
(3) 序列的重排和缺失；
(4) 重复序列污染：cDNA 克隆方法有时会受到逆转录因子(如 Alus)的影响。
(5) 测序误差和自然多态性：测序过程存在一定的误差概率。

对付以上这些偏差，一个聪明的策略是用可能污染数据记录的序列(如载体)去估计误差程度。同时，一些去除污染的专门软件系统已被研制出来，如 EBI 网站便提供了去除载体污染的在线服务，网址为 http://www.ebi.ac.uk/blastall/vectors.html。EMBL 研制了基于 BLAST 的载体扫描服务和一个特殊的序列数据库 EMVEC。EMVEC 的序列来自 EMBL 的 SYN(synthetic division)类 2000 余条一般用于克隆和测序实验的序列，该库随着 EMBL 的扩充而实时更新。

[1]本部分内容主要取自F．奥斯伯，R．E．金斯顿等．精编分子生物学实验指南，北京：科学出版社，1998

第五节 向数据库发送序列数据及其它[2]

本节将简单介绍如何向相关数据库发送自己的序列数据，如何准确、全面表述生物信息学研究的“材料与方法”和普通用户可利用的数据库服务内容。

许多学术期刊在发表含有序列数据的论文时，均要求作者先将该序列发送并存贮到某一数据库中。如果该序列是在欧洲完成的，则应储存到 EMBL，如来自美洲，则存到 GenBank，其它地区则应发送到日本的 DDBJ。这些数据库的主页上均有详细的发送说明。数据库往往特别要求发送者要注意去除载体污染，例如 EMBL 提供了 EBI 的相关服务(网址见上节)。序列的发送可以通过网上进行。EMBL 的发送系统为 WEBIN(http://www.ebi.ac.uk/embl/Submission/webin.html)，它除了可进行一般大小的序列数据发送外，还可进行大批量的数据发送(Bulk submission)。GenBank 的发送系统 Sequin(http://www.ncbi.nlm.nih.gov/Sequin/index.html)是由 NCB1 开发的多平台(Mac/pc/unix)工具，适用于 EMBL、GenBank 和 DDBJ 数据库的发送服务。具体发送格式和要求可到这些网站上查获。一旦数据被接收，一个记录号(对应于发送的数据)将产生并送给发送者，该记录号可用于论文发表。但发送的序列在公共数据库中出现可能会有一个滞后期，因为注释和核查将颇费一番周折。

试验结果的可重复性是科学研究的一个重要特征。为了保证生物信息学研究结果的可重复性，准确、全面的“材料与方法”说明比其它学科显得更为重要和严格。一份清楚、准确的“材料与方法”说明应包括：

(1)数据库的名称：SWISSPROT、PIR、GenBank、EMBL、dbEST 等等，不应是以类别(蛋白、核酸、序列等)说明。

(2)数据库的版本(Version)：数据库在快速变化，它远快于期刊的发行速度，所以严格注明所用数据库的版本；如果你的检索是实时的，则注明最后检索的日期。

(3)所使用的计算机：这可能是不重要的一项说明，因为算法等不论在何种计算机上均应相同，但如果在使用异地(off-site)计算机系统(如 E-mail 和 Internet)服务，那么，科学的态度应是注明其服务器及其管理者。

如果进行序列的比较研究，还应包括以下内容（具体内容可参见第三章第 2 节）：

(4)替换矩阵(substitution matrix)：所有的现代搜索程序均使用替换矩阵，选用不同的矩阵会产生完全不同的结果，所以必须注明在搜索和列阵(aligning)中使用何种矩阵。

(5)空位罚值(gap penalty)：很多算法使用空位罚值(如 FASTA)。

一般用户可利用的分子数据库服务内容可分为几种：E-mail 服务、匿名 FTP 服务、www 服务和序列相似性搜索服务等。通过 E-mail 可向数据库发送相关要求来获取有关数据和服务。例如，可发一个服务指令到 EBI 的 mail to：netserv.ebi.ac.uk 地址，服务指令中应以一个指令开头，比如你想获得记录号为 X55652 的 DNA 序列，则应在指令栏中键入“GET NUC：X55652”，这样 EBI 服

[2]本部分内容主要取自F．奥斯伯，R．E．金斯顿等．精编分子生物学实验指南，北京：科学出版社，1998

务器便会将该序列的信息发到你的信箱中。匿名 FTP 服务是另外一种进入数据库获取信息的方法，研究者可利用本地的 FTP(file transfer protocol)程序连接到相应的数据库主机上，以“ anonymous ”(匿名)为用户名和自己的 E-mail 地址为口令进入。www 服务是通过网络直接进入相关数据库网址，进行数据检索、数据传送等。同时各数据库均提供序列相似性检索等序列分析的服务，如 FASTA、BLAST 和 BLITS 等服务(具体说明见第三章第 3 节)，分析结果通过 E-mail 发送返回或直接显示在浏览器上。

第三章 序列分析与联配

序列分析是生物信息学最主要的研究内容之一，它可以分为两个主要部分：一是序列组成（特别是涉及到基因组层次上）分析，二是序列之间的比较分析。两条序列或多条序列间的比对或联配(alignment)的目的，是对它们的序列相似性进行评估，找出这些序列中结构或功能相似性区域等。通过联配未知序列与已知序列(其功能或结构等已知)的相似程度，我们可以判断或推测未知序列的结构与功能。

第一节　序列组成及单一序列分析[1]

一．碱基组成

DNA 序列一个显而易见的特征是四种碱基类型的分布。尽管四种碱基的频率相等时对数学模型的建立可能是方便的，但几乎所有的研究都证明碱基是以不同频率分布的。表 3.1 包含了 9 条完整 DNA 分子序列的资料，表 3.2 的数据来自两个胎儿球蛋白基因(Gr 和 Ar)，每个基因具有三个外显子和两个内含子(shen 等 1981)。这两个例子说明序列内和序列间碱基具有不同的频率。在基因每一侧的 500 个任意碱基区域被称为“侧翼”，基因间区域是指两个基因间的其余序列。

表 3.1 九条完整 DNA 序列的碱基组成

序　列	名　称*	碱基频率				总　计
		A	C	G	T	
噬菌体						
λ	LAMCG	0.25	0.24	0.25	0.26	48502
T_7	PT7	0.27	0.23	0.24	0.26	39936
ØX174	PX1CG	0.24	0.22	0.31	0.23	5386
病毒						
花椰菜镶病毒	MCACGDH	0.37	0.21	0.23	0.19	8016
人类乳头多瘤空泡病毒 BK	PVBMM	0.30	0.20	0.30	0.20	4936
肝炎 B	HPBAYW	0.28	0.22	0.23	0.27	3182
线粒体						
人类	HUMMT	0.31	0.31	0.25	0.13	16569
牛	BOVMT	0.33	0.26	0.27	0.14	16338
鼠	MUSMT	0.35	0.24	0.29	0.12	16295

*取自 GenBank 数据库

[1]部分内容取自Weir B. S.（徐云碧等译）．遗传学数据分析—群体遗传学离散型数据分析方法，北京：中国农业出版社，1996

表 3.2 人类胎儿球蛋白基因不同区段的碱基组成

区 段	长 度	A	C	G	T
5′例翼(2)	1000	0.33	0.23	0.22	0.22
3′例翼(2)	1000	0.29	0.15	0.26	0.30
内含子(4)	1996	0.27	0.17	0.27	0.29
外显子(6)	882	0.24	0.25	0.28	0.22
基因间(1)	2487	0.32	0.19	0.18	0.31

*数据来自 EMBL 数据库 HSGLBN 基因

二. 碱基相邻频率

分析 DNA 序列的主要困难之一是碱基相邻的频率不是独立的。碱基相邻的频率一般不等于单个碱基频率的乘积：如果 P_u 是序列中碱基 u 的频率，且 P_{uv} 为两个相邻碱基 u 和 v 的频率，则

$$p_{uv} \neq p_u p_v$$

Nussinov(1984)研究了两碱基相邻的频率(表 3.3)。数据来自 166 个脊椎动物的 DNA 序列，总长 136731 个碱基。表中的比值为 16 种二个碱基相邻的频率除以相应的单个碱基频率的乘积。

表 3.3 脊椎动物中两碱基的相邻频率

相邻碱基对	观测频率/期望频率*
TG	1.29
CT	1.26
CC	1.18
AG	1.16
AA	1.15
CA	1.15
GG	1.14
TT	1.07
GA	10.4
TC	1.00
GC	0.99
AT	0.85
AC	0.84
GT	0.82
TA	0.65
CG	0.42

*期望频率为相应两个单个碱基频率的乘积

作为一个特别的例子，图 3.1 给出了鸡血红蛋白 β 链的 mRNA 编码区的 438 个碱基。表 3.4 列出了 4 种碱基和 16 种两碱基的数目。将该表看作 4×4 的表，

计算行列独立性的卡方统计量，得到 $\chi^2 = 59.3(\chi^2_{0.05,9} = 16.92)$，表明行(第一碱基)列(第二碱基)之间存在明显的关联。

```
GTGCACTGGA  CTGCTGAGGA  GAAGCAGCTC  ATCACCGGCC  TCTGGGCAA   GGTCAATGTG  60
GCCGAATGTG  GGGCCGAAGC  CCTGGCCAGG  CTGCTGATCG  TCTACCCCTG  GACCCAGAGG  120
TTCTTTGCGT  CCTTTGGGAA  CCTCTCCAGC  CCCACTGCCA  TCCTTGGCAA  CCCCATGGTC  180
CGCGCCCACG  GCAAGAAAGT  GCTCACCTCC  TTTGGGGATG  CTGTGAAGAA  CCTGGACAAC  240
ATCAAGAACA  CCTTCTCCCA  ACTGTCCGAA  CTGCATTGTG  ACAAGCTGCA  TGTGGACCCC  300
GAGAACTTCA  GGCTCCTGGG  TGACATCCTC  ATCATTGTCC  TGGCCGCCCA  CTTCAGCAAG  360
GACTTCACTC  CTGAATGCCA  GGCTGCCTGG  CAGAAGCTGG  TCCGCGTGGT  GGCCCATGCC  420
CTGGCTCGCA  AGTACCAC
```

图 3.1 鸡 β 球蛋白基因编码区的 DNA 序列
(GenBank：CHKHBBM，记录号 J00860)

表 3.4 图 3.1 鸡 β 球蛋白基因序列的相邻碱基分布

		第二碱基				总计
		A	C	G	T	
第一碱基	A	23	26	23	15	87
	C	37	51	14	41	143
	G	25	38	36	19	118
	T	2	29	41	14	89
总计		87	144	117	89	437

在编码区，存在某种约束来限制 DNA 序列编码氨基酸。在密码子水平上，这一约束与碱基相邻频率有关。表 3.5 列出了遗传密码和图 3.1 序列中各密码子数量。尽管数目很小，难以作出有力的统计结论，但编码同一氨基酸的不同密码子(同义密码子)好像不是等同存在的。这种密码子偏倚必定与两碱基相邻频率水平有关。表 3.5 还清楚地表明，由于密码子第 3 位置上碱基的改变常常不会改变氨基酸的类型，因而对第 3 位置上碱基的约束要比第 2 位碱基小得多。

表 3.5 64 种可能的碱基三联体密码子及相应的氨基酸数（据图 3.1 序列）

UUU Phe 3	UCU Ser 0	UAU Tyr 0	UGU Cys 2
UUC Phe 5	UCC Ser 5	UAC Tyr 2	UGC Cys 1
UUA Leu 0	UCA Ser 0	UAA Stop 0	UGA Stop 0
UUG Leu 0	UCG Ser 0	UAG Stop 0	UGG Trp 4
CUU Leu 1	CCU Pro 1	CAU His 3	CGU Arg 0
CUC Leu 6	CCC Pro 4	CAC His 4	CGC Arg 3
CUA Leu 0	CCA Pro 0	CAA Gln 1	CGA Arg 0
CUG Leu 11	CCG Pro 0	CAG Gln 0	CGG Arg 0
AUU Ile 1	ACU Thr 3	AAU Asn 1	AGU Sre 0
AUC Ile 6	ACC Thr 4	AAC Asn 6	AGC Ser 2
AUA Ile 0	ACA Thr 0	AAA Lys 1	AGA Arg 0
AUG Met 1	ACG Thr 0	AAG Lys 9	AGG Arg 3
GUU Val 0	GCU Ala 4	GAU Asp 1	GGU Gly 1
GUC Val 5	GCC Ala 11	GAC Asp 5	GGC Gly 4
GUA Val 0	GCA Ala 0	GAA Glu 4	GGA Gly o
GUG Val7	GCG Ala 1	GAG Glu 3	GGG Gly 3

相邻碱基之间的关联将导致更远碱基之间的关联，这些关联延伸距离的估计可以从马尔科夫链(Markov chain)理论得到(Javare 和 Giddings，1989)。在不援引任何生物学机制的情况下，第 k 阶马尔科夫链假定在序列中某一位置上碱基的存在只取决于前面 k 个位置上的碱基。一阶链假定一个特定碱基存在于位置 i 的概率只取取决于在位置 i-1 的 4 种碱基概率。相互独立的碱基所组成的序列将与 0 阶马尔科夫链相对应。阶可以通过似然法估计。同时，马尔科夫链分析更适应于基因组水平，而非单一序列(基因)。相关内容可参见第四章第 2 节。

三．同向重复序列分析

除了分析整个序列碱基关联程度的特征外，我们常对寻找同向重复序列(direct repeats)之类的问题感兴趣。Karlin等(1983)给出了完成这一分析的有效算法。该法采用由特定的几组碱基字母组成的不同亚序列或称为字码(word)。只需要对整个序列搜索一次。给一碱基赋以值α，例如A、C、G、T的值为 0、1、2、3。由X_1、X_2、…、X_k共k个字母组成的每一种不同的字码按$1+\sum_{i=1}^{k}\alpha_i 4^{k-i}$ 计算字码值。这些值的取值范围为 1 到 4^k。例如，5 字码TGACC的值为 $1+3\times4^4+2\times4^3+0\times4^2+1\times4^1+1\times4^0=459$。可先从低k值的字码开始搜索。记录序列中每一个位置k字码的字码值。只有在发现k字码长度重复的那些位置考虑进行长度大于k的字码搜索。

表 3.6 列出了序列 TGGAAATAAAACGTAAGTAG 中所有碱基 2 字码(k=2)的初始位置和字码值。对于完全重复、长度大于 2 的同向重复或亚序列的搜索可只限于 2 字码重复的初始位置。在本例中只有 4 个重复的 2 碱基重复序列。例如，在位置 4、5、8、9、10 和 15 均发现了字码值为 1 的碱基重复序列。从有重复的第 2 个碱基为起点的 3 字码值及位置列于表 3.7，其中发现字码值为 1、45 和 49 的序列有重复。以每一重复的 3 碱基为起点的 4 字码搜索未能发现更长的重复序列。

因此最长的同向重复为 4、8、9 位置上的 AAA，13、17 位置上的 GTA 以及 7、14 位置上的 TAA。同样对图 3.1 鸡β球蛋白 DNA 序列进行同向重复序列搜索，一些最长同向重复序列列于表 3.8。

表 3.6 序列 TGGAAATAAAACGTAAGTAAGTAG 的 2 字码值和位置(Karlin，1983)

字码值	碱基位置	字码值	碱基位置
1	4,5,8,9,10,15	9	3
2	11	10	-
3	16,19	11	2
4	6	12	13,17
5	-	13	7,14,18
6	-	14	-
7	12	15	1
8	-	16	1

表 3.7 序列 TGGAAATAAAACGTAAGTAG 的 3 字码值和位置(Karlin, 1983)

字码值	碱基位置
1	4,8,9
2	10
3	15
4	5
45	13,17
49	7,14
51	18

表 3.8 鸡β球蛋白 DNA 序列中(图 3.1)长度为 8 或 8 以上的碱基重复序列

长度	重复序列	起始位置
8	GCCCTGGC	79，418
	GCCAGGCT	85，377
	CCAGGCTG	86，378
	CAGGCTGC	87，379
	TCCTTTGG	130，208
	CCTTTGGG	131，209
	TGGTCCGC	176，398
	GGTCCGCG	177，399
9	GCCAGGCTG	85，377
	CCAGGCTGC	86，378
	TCCTTTGGG	130，208
	TGGTCCGCG	176，398
10	GCCAGGCTGC	85，377

Karlin等(1983)提出了序列内存在的最长同向重复序列的统计显著性评价

方法。在核苷酸的位置为独立的假定下(相当于阶次为 0 的马尔科夫链)，长度为n的序列中，最长同向重复*Ln*的期望长度和方差为：

$$\mu_L = \frac{0.6359 + 2\ln n + \ln(1-p)}{\ln(1/p)} - 1$$

$$\sigma_L^2 = \frac{1.645}{(\ln P)^2} \qquad (3.1)$$

其中，*P*为序列中碱基频率的平方和：

$$P = \sum_{i=1}^{4} P_i^2$$

用尽可能接近最大长度的期望均值的字码(即R→ μ_L) 来开始同向重复序列的搜索计算可能节省计算量。

可以用一个近似方法来验证以上统计假说。假定同向重复序列的长度呈正态分布。对于图 3.1 鸡β蛋白序列，A、C、G、T 四个碱基的次数分别为 87、144、118 和 89，因而 P=0.2614，最长重复序列的期望长度为 8.13 且具有期望方差 0.9138。根据 95%的正态分布概率，理论上可以预期最长同向重复序列不超过 10。

四． DNA 序列的几何学分析—Z 曲线

DNA 序列实际上是一种用 4 种字母表达的“语言”，只是其“词法”和“语法”规则目前还没有搞清楚。人类的语言有文字、声音两种基本表现形式，此外还有手语、旗语甚至图画语等特殊表达形式。同样，DNA 序列作为一种语言，其表达形式也不是唯一的。传统上，DNA 序列是用 4 种字母符号表达的一维序列。这是一种抽象形式，适合于存储、印刷和代数算法的处理，包括比较、排列和查找特殊序列等。我国学者张春霆等开展了 DNA 序列三维空间曲线表示形式，即 DNA 序列几何表示形式的研究。几何形式虽然与符号形式完全等价，但显示了 DNA 序列的新特征。两种形式各有其特点，相互补充。这一新方法，为解读 DNA 序列信息提供了崭新的手段。

他们的研究始于对 4 种碱基对称性的观察，提出了用正面体表示碱基对称性。1994 年，他们利用这种形式来表示任意长度的 DNA 序列。现将这种序列表示方法简述如下。

考察一个长为L的单股DNA序列，方向(5'→3'或 3'→5')不限。从第一个碱基开始，依次考察此序列，每次只考察一个碱基。当考察到第n个碱基时(n=1,2，…,L)，数一下从 1 到n这个子序列中四种碱基各自出现的次数。设 4 种碱基A、C、G、T出现的次数分别以A_n、C_n、G_n、T_n表示之，这里下标“n”是表明这些整数是从 1 到n这个子序列中数出来的，如图 3.2 所示。显然，它们都是正整数。根据正四面体的对称性可以证明，在正面体内存在唯一的一个点P_n与这四个正整数对应。点P_n构成了四个正整数的一一对应映射。点P_n坐标可用四正整数表达：

$$x_n=2(An+Gn)-n,$$
$$y_n=2(An+Gn)-n, \qquad (3.2)$$
$$z_n=2(An+Tn)-n,$$

x_n，y_n，z_n∈[-n，n]，n=1，2，…，L，

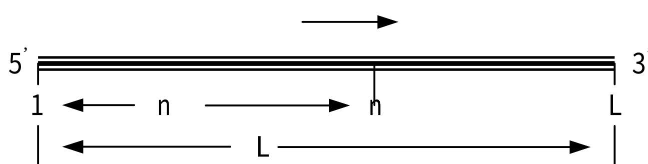


图 3.2 DNA 序列示意图

其中x_n，y_n和z_n为点P_n的三个坐标分量。当n从 1 跑到L时，我们依次得到P_1，P_2，P_3，…，P_L共L个点。将相邻两点用适当的曲线连接所得到的整条曲线，就称为表示DNA序列的Z曲线。可以证明，Z曲线与所表示的DNA序列是一一对应的，即给定一DNA序列，存在唯一的一条Z曲线与之对应；反之，给定一条Z曲线，可找到唯一的一个DNA序列与之对应。换言之，Z曲线包含了DNA序列的全部信息。Z曲线是与符号DNA序列等价的另一种表示形式，一种几何形式。可以通过Z曲线对DNA序列进行研究。

Z曲线的三个分量(方程 3.2)具有明确的生物学意义：x_n表示嘌呤/嘧啶碱基沿序列的分布。当从 1 到n的这个子序列中(图 3.2)嘌呤碱基多于嘧啶碱基时，$x_n>0$，否则，$x_n<0$，当两者相等时$x_n=0$。同样，y_n表示氨基/酮基碱基沿序列的分布。当在子序列中氨基碱基多于酮基碱基时，$y_n>0$，否则，$y_n<0$，当两者相等时$y_n=0$。z_n表示强/弱氢键碱基沿序列的分布。当弱氢键碱基多于强氢键碱基时，$z_n>0$，否则$z_n<0$，当两者相等时，$z_n=0$。这三种分布是相互独立的，表现在以下事实上：任何一种分布不能由其它两种分布的线性叠加表示出来。给定的DNA序列唯一地决定了这三种分布；三种分布唯一地描述了DNA序列。对DNA序列的研究就是通过对这三种分布的研究来进行。从方法学的角度来看，这是DNA序列的一种几何学研究途径。

图 3.3 给出了大肠杆菌 ayoP 基因族序列 Z 曲线的三个分量，即三种分布图。该基因族包含了大肠杆菌 5 个基因 aroP，A，aceFE，aceF 和 lpd，总长度为 9501bp，分别编码芳香族氨基酸运输蛋白 aroP，蛋白质 A(功能不详)和三种酶，即丙酮酸脱氢酶，二氢硫辛酰基转移酶和二氢硫辛酰脱氢酶。它们位于此序列的 0039-1406，1947-2654，2870-5527，5545-7434，7759-9183 区间。在图中 X 轴的下方的基因排列图上已分别用阴影标出相应基因。在这些基因之间有三个启动子区（*pm1*、*pm2* 和 *pm3*），其中 aceE 和 aceF 基因属于 ace 操纵子，共用一个启动子。三个启动子区亦在图中标出。非常令人感兴趣的是，在 5 个编码区，Z 曲线的 z 分量基本上都是单调下降的，而在三个启动子区基本上都是单调上升的。x，y 分量亦有变化，但不如 z 分量明显。在上升、下降的交界处，Z 曲线均发生了重大的转折，据此有可能用 Z 曲线识别这些位置。由此图可见，用 Z 曲线这种几何方法显示 DNA 序列不仅直观，而且作为一种识别序列中的不同基因和功能区的新方法，展现了广阔的应用前景。

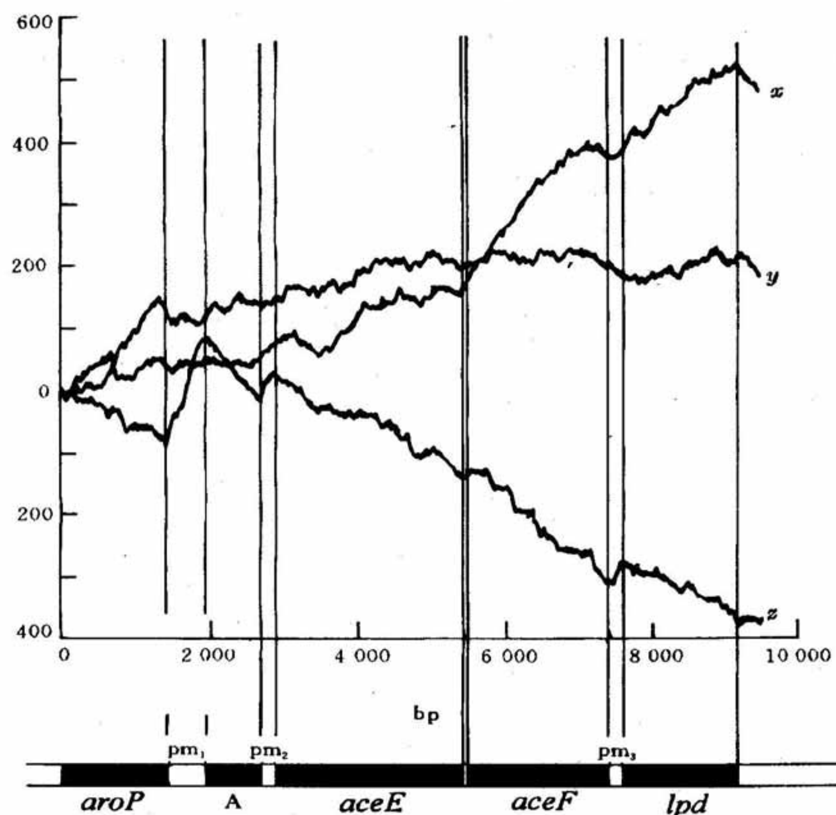


图 3.3 大肠杆菌 ayoP 基因族序列 Z 曲线的三个分量（三种分布图）

第二节 序列联配[2]

一．Needleman-Wunsch 算法

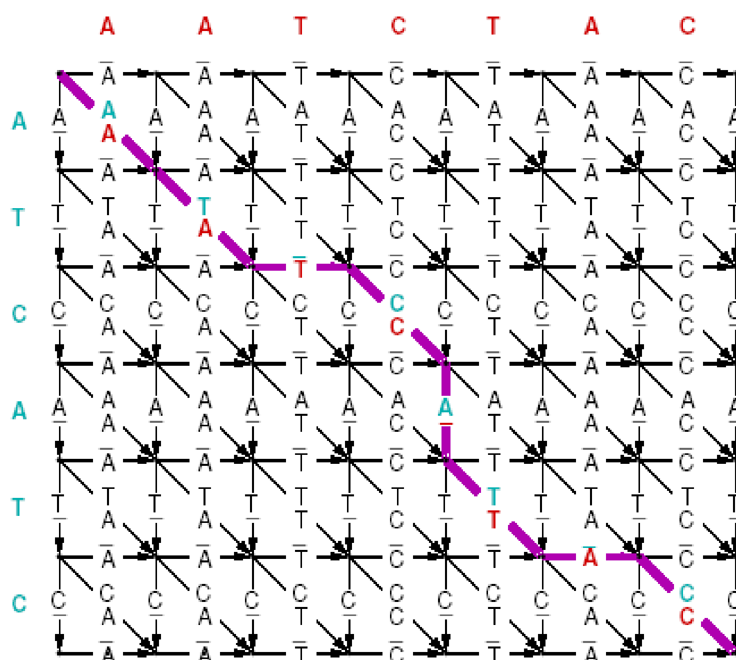
有 2 种经典方法可以计算两条序列间的最适联配。Needleman-Wunsch 算法是一种整体联配(global alignment)算法，最佳联配中包括了全部的最短匹配序列。Smith-Wateman 算法是在 Needleman-Wunsch 算法基础上发展而来的，它是一种局部联配(Local alignment)算法。这二种算法均可以用于核酸和蛋白质序列。在给定空位罚值和替换矩阵情况下，它们总是能给出具有最高（优）联配值的联配。但是，这个联配并不需要达到生物学意义上的显著水平。GCG 软件包中，BESFIT 和 GAP 程序，EMBOSS 的 needle 等可用于该联配。一些网站可以通过递交序列进行两条序列的联配分析。

从整体上分析两个序列的关系，即考虑序列总长的整体比较，用类似于使整体相似 (global similarity)最大化的方式，对序列进行联配。两个不等长度序列的联配分析必需考虑在一个序列中圈掉一些碱基或在另一序列作空位(gap)处理。Needleman 和 Wunsch(1970)的法则为这些步骤提供了实例。这一算法是为氨基酸序列发展的，但也可以用于核苷酸序列。算法最初寻求的是使两条序列间

[2]部分内容取自Weir B. S. Genetic Data AnalysisⅡ—Methods for Discrete Population Genetic Data， Sunderland：Sinauer Associates Inc. Publishes，1996

的距离最小。尽管这类距离的元素是以一种特定的方式定义的，但该算法的良好特性在于它确定了最短距离。这是一个动态规划(dynamic programming)的方法。

将两条联配的序列沿双向表的轴放置，两条序列的所有可能的联配方式都将在它们所形成的方形图中（见下图）。从任一碱基对，即表中的任一单元开始，联配可延三种可能的方式延伸：如果碱基不匹配，则每一序列加上一个碱基，并给其增加一个规定的距离权重；或在一个序列中增加一个碱基而在另一序列中增加一个空位或反之亦然。引入一个空位时也将增加一个规定的距离权重。因此，表中的一个单元可以从(至多)三个相邻的单元达到。我们把达左上角单元距离最小的方向看作相似序列延伸的方向。等距离时意味着存在两种可能的方向。将这些方向记录下来，并在研究了所有的单元之后，沿着记录的方向就有一条路径可从右下角(两个序列的末端)追踪到左上角（两个序列的起点）。由此所产生的路径将给出具有最短距离的序列联配。



Alignment corresponding to the colored path:

A T – C A T – C
A A T C – T A C

以两个短序列 CTGTATC 和 CTATAATCCC 为例，将上述过程说明于图 3.4。设碱基错配时距离权重为 1，引入一个空位时距离权重为 3。该图边缘的行和列作为起始条件增加到表中。在单元 5 行 3 列，即相应较短序列(第二序列)的第 2 个 T 碱基和较长序列(第一序列)的第 1 个 T 碱基位置，有三种可能的距离增量。设在各序列中增加碱基 T 时 （从 4 行 2 列移动)对距离的贡献为 0。从 5 行 2 列的位置作水平移动(等价于增加第二序列的碱基 T 而在第一序列引入一个空位)，在本例中增加一个罚值 3。从 3 列 4 行向该单元作垂直移动，使第一序列增加碱基 T 而第二序列引入一个空位，结果也得到一个罚值 3。因此从该单元(5 行 3

列)所得到的最小距离的延伸方向是沿对角线和水平方向。在表中这两个方向用箭头表示。这两种最短方向都使从左上角到该单元的距离为 6。沿箭头所指方向在表中从右下角向左上角追踪，得到 6 种可能的联配：

```
CTATAATCCC        CTATAATCCC        CTATAATCCC
CTGTA-TC--        CTGTA-T-C-        CTGTA-T--C

CTATAATCCC        CTATAATCCC        CTATAATCCC
CTGT-ATC--        CTGT-AT-C-        CTGT-AT--C
```

在上述 6 种联配中，距离均为 10，即在较短序列中有 6 个匹配碱基、1 个错配碱基和 3 个空位。

	O			C			T			A			T			A			A			T			C			C			C
O	0	←	3	3	←	3	6	←	3	9	←	3	12	←	3	15	←	3	18	←	3	21	←	3	24	←	3	27	←	3	30
	↑			↑																											
C	3		0	3		1	3		1	3		1	3		1	3		1	3		1	3		0	3		0	3		0	3
	3		3	0	←	3	3	←	3	6	←	3	9	←	3	12	←	3	15	←	3	18	←	3	21	←	3	24	←	3	27
	↑			↑																											
T	3		1	3		0	3		1	3		0	3		1	3		1	3		0	3		1	3		1	3		1	3
	6		3	3		3	0	←	3	3	←	3	6	←	3	9	←	3	12	←	3	15	←	3	18	←	3	21	←	3	24
	↑			↑			↑							←																	
G	3		1	3		1	3		1	3		1	3		1	3		1	3		1	3		1	3		1	3		1	3
	9		3	6		3	3		3	1	←	3	4	←	3	7	←	3	10	←	3	13	←	3	16	←	3	19	←	3	22
	↑			↑			↑			↑																					
T	3		1	3		0	3		1	3		0	3		1	3		1	3		0	3		1	3		1	3		1	3
	12		3	9		3	6		3	4		3	1	←	3	4	←	3	7	←	3	10	←	3	13	←	3	16	←	3	19
	↑			↑			↑						↑																		
A	3		1	3		1	3		0	3		1	3		0	3		0	3		1	3		1	3		1	3		1	3
	15		3	12		3	9		3	6		3	4		3	1	←	3	4	←	3	7	←	3	10	←	3	13	←	3	16
	↑			↑			↑			↑						↑															
T	3		1	3		0	3		1	3		0	3		1	3		1	3		0	3		1	3		1	3		1	3
	18		3	15		3	12		3	9		3	6		3	4		3	2		3	4	←	3	7	←	3	10	←	3	13
	↑			↑			↑			↑			↑			↑			↑												
C	3		0	3		1	3		1	3		1	3		1	3		1	3		1	3		0	3		0	3		0	3
	21		3	18		3	15		3	12		3	9		3	7		3	5		3	3		3	4	←	3	7	←	3	10

图 3.4 Needleman-Wusch 算法实例。设定碱基错配的距离权重为 1，单个碱基缺失或插入时距离权重为 3

该算法可以用代数形式来描述。设具有碱基a_i和b_j的两个序列a和b，这两个序列间距离为d(a,b)。通过评价序列a中前i个位置和序列b前j位置的距离$d(a^i,b^j)$，递归地得到距离d(a,b)。如果a和b的长度为m和n，则其期望距离为$d(a^m,b^n)$。上表中引入的第1行1列单元的距离为0(相当于空序列),在单元(i,j)内，使到达该单元距离增加的三种可能事件为：

1.从单元(i-1,j)向(i,j)的垂直移动，相当于在b序列中插入一个空位使相似序列延伸。换言之，b序列由a序列中a_i的缺失所产生，这一事件的权重记作$w_(a_i)$。

2.从单元(i-1,j-1)向(i,j)的对角线移动，相当于增加碱基a_i和b_j使相似序列延伸。换言之，b序列由a序列中的a_i被b_j取代所产生，这一事件的权重记为$w_(a_i,b_j)$。

3.从单元(i,j-1)向(i,j)的水平移动，相当于在序列b中插入一个空位使相似序列延伸。换言之，b序列由b_j插入a序列所产生，这一事件的权重记为$w_+(b_j)$。

因此，单元(i,j)的距离$d(a^i,b^j)$可看成三个相邻单元的距离加上相应权重后的最小者，即

$$d(a^i,b^j)=\min\begin{cases}d(a^{i-1},b^j)+w_-(a_i)\\ d(a^{i-1},b^{j-1})+w(a_i,b_j)\\ d(a^i,b^{j-1})+w_+(b_i)\end{cases} \qquad (3.3)$$

且初始条件为

$$d(a^0,b^0)=0$$

$$d(a^0,b^j)=\sum_{k=1}^{j}w_+(b_k)$$

$$d(a^i,b^0)=\sum_{k=1}^{i}w_-(a_k)$$

在图3.4的实例中

$$w_-(a_i)=3 \qquad \text{(对于每一个 i)}$$

$$w(a_i,b_j)=\begin{cases}0 & (i=j)\\ 1 & (i\neq j)\end{cases}$$

$$w_+(b_j)=3 \qquad \text{(对于每一个 j)}$$

当两个序列被联配时，通过计算其重排序列(shuffed version)的联配距离，可以得到这两个序列间的最小距离估计。如果实际得到的联配距离小于重排序列距离的95%，则表明实际的联配距离达到了5%的显著水平，是不可能由机误造成的。

二．Smith-Waterman 算法

由于亲缘关系较远的蛋白质序列可能只有一些相互独立的相同片段，所以进行局部相似性分析有时可能比整体相似性分析更合理。Smith和Waterman描述了一种查找具有最高相似性片段的算法。对于序列A=($a_1,a_2,\ldots,a_m$)和B=($b_1,b_2,\ldots,b_n$)，H_{ij}被定义为以a_i和b_j碱基对结束的片段(亚序列)的相似性值。与Needle-Wunsch算法一样，Smith-Waterman算法也要利用递推关系来确定H值，H的初始值为：

$$H_{i0}=0，\quad 0\le i\le n，\quad H_{0j}=0，\quad 0\le j\le m$$

相似性计算中包括 2 个统计量：碱基对(序列因子) a_i,b_j 的相似性值 $S(a_i,b_j)$ 和空位权重 $w_k=v+uk$ (k 为空位长度)。Smith-Waterman 算法可以给出 2 条序列的最大相似性值。以 a_i,b_j 碱基对结束的片段可以由以 a_{i-1} 和 b_{j-1} 结束片段增加碱基(因子)来获得，或者 a_i 可以删除 k 长度的碱基片段，b_j 可删除 l 长度碱基片段。具体算法如下：

$$\begin{aligned}P_{ij}&=\max(H_{i-1,j}-w_1,P_{i-1,j}-u)\\Q_{ij}&=\max(H_{i,j-1}-w_1,P_{i,j-1}-u)\end{aligned}\qquad(3.4)$$

则 $$H_{ij}=\max\begin{cases}H_{i-1,j-1}+S(a_i,b_j)\\P_{ij}=\max_{1\le k\le i}(H_{i-k,j}-w_k)\\Q_{ij}=\max_{1\le l\le j}(H_{i,j-l}-w_l)\\0\end{cases},(1\le i\le m,1\le j\le n)\qquad(3.5)$$

其中 $P_{0,0}=P_{0,j}=Q_{0,0}=Q_{i,0}=0$

该算法可以确保具有最大H_{ij}值的序列片段是相似性最好的。从(a_i,b_j)为起点，向后追踪H_{ij}矩阵，直到到达某一负值。对于具有最大相似性片段以外部分的差异性不会影响到该片段的H值。

举例说明了这一算法。我们同样以上节 Needleman-Wunsch 算法中的两条短序列为例。两条序列(CTGTATC 和 CTATAATCCC)排于表 3.9 的两侧，相应的H_{ij}，P_{ij}和Q_{ij}值分别列入表中。本例的权重等根据 Smith 和 Waterman(1981)以前的例子设定为：

$$S(a_i,b_j)=\begin{cases}1 & a_i=b_j\\ -1/3 & a_i\neq b_j\end{cases}$$

$$w_k=1+k/3 \qquad (3.6)$$

对于 4 个碱基具有相同频率的随机长序列，$S(a_i,b_j)$值的平均值为零。w_k值应至少不小于匹配与不匹配权重的差值。

表 3.9 的最大H_{ij}为 4.33(8 行与 7 列相交处)，星号(*)表示出具有最大相似性的片段匹配方式：

```
CTGTA-TC
CTATAATC
```

表 3.9 Smith-Waterman 算法例举

			j=0	j=1	j=2	j=3	j=4	j=5	j=6	j=7
			0	C	T	G	T	A	T	G
i=0	0	H_{ij}	0	0	0	0	0	0	0	0
		P_{ij}	0	0	0	0	0	0	0	0
		Q_{ij}	0	0	0	0	0	0	0	0
i=1	C	H_{ij}	0	1.00*	0.00	0.00	0.00	0.00	0.00	1.00
		P_{ij}	0	-0.33	-0.33	-0.33	-0.33	-0.33	-0.33	-0.33
		Q_{ij}	0	-0.33	-0.33	-0.67	-1.00	-1.33	-1.33	-1.33
i=2	T	H_{ij}	0	0.00	2.00*	0.67	1.00	0.00	1.00	0.00
		P_{ij}	0	-0.33	-0.67	-0.67	-0.67	-0.67	-0.67	-0.33
		Q_{ij}	0	-0.33	-0.67	0.67	0.33	0.00	-0.33	-0.33
i=3	A	H_{ij}	0	0.00	0.67	1.67*	0.33	2.00	0.67	0.67
		P_{ij}	0	-0.67	0.67	-0.67	-0.33	-1.00	-0.33	-0.67
		Q_{ij}	0	-0.33	-0.67	-0.67	0.33	0.00	0.67	0.33
i=4	T	H_{ij}	0	0.00	1.00	0.33	2.67*	1.33	3.00	1.67
		P_{ij}	0	-1.00	0.33	0.33	-0.67	0.67	-0.67	-0.67
		Q_{ij}	0	-0.33	-0.67	-0.33	-0.67	1.33	1.00	1.67
i=5	A	H_{ij}	0	0.00	0.00	0.67	1.33	3.67*	2.33	2.67
		P_{ij}	0	-1.33	0.00	0.00	1.33	0.00	1.67	0.33
		Q_{ij}	0	-0.33	-0.67	-1.00	-0.67	0.00	2.33	2.00
i=6	A	H_{ij}	0	0.00	0.00	0.00	1.00	2.33*	3.33	2.00
		P_{ij}	0	-1.33	-0.33	-0.33	1.00	2.33	1.33	1.33
		Q_{ij}	0	-0.33	-0.67	-1.00	-1.33	-0.33	1.00	2.00
i=7	T	H_{ij}	0	0.00	1.00	0.00	1.00	2.00	3.33*	3.00
		P_{ij}	0	-1.33	-0.67	-0.67	0.67	2.00	2.00	1.00
		Q_{ij}	0	-0.33	-0.67	-0.33	-0.67	-0.33	0.67	2.00
i=8	C	H_{ij}	0	1.00	0.00	0.67	0.33	1.67	2.00	4.33*
		P_{ij}	0	1.33	-0.33	-1.00	0.33	1.67	2.00	1. 67
		Q_{ij}	0	-0.33	-0.33	-0.67	-0.67	1.00	0.33	0. 67
i=9	C	H_{ij}	0	1.00	0.67	0.00	0.33	1.33	1.67	3. 00
		P_{ij}	0	-0.33	-0.67	-0.67	0.00	1.33	1.67	3. 00
		Q_{ij}	0	-0.33	-0.33	-0.67	-1.00	-1.00	0.00	0. 33
i=10	C	H_{ij}	0	1.00	0.67	0.33	0.00	1.00	1.33	2. 67
		P_{ij}	0	-0.33	-0.67	-1.00	-0.33	1.00	1.33	2. 67
		Q_{ij}	0	-0.33	-0.33	-0.67	-1.00	-1.33	-0.33	0. 00

三．序列相似性的统计特性[3]

到目前为止，对局部联配的统计学问题已基本搞清楚，特别是那些不含有空位(gap)的局部联配更是如此。我们不妨首先考虑不含有空位的局部联配问题，BLAST 最初的搜索程序便是以此为基础的。

无空位局部联配涉及的是等长度的一对序列片段，两个片段的各部分彼此比较。一种 Smith-Waterman 或 Sellers 算法的改进算法可以找到所有高比值片段对(high-scoring segment pairs,HSPs)，即这些片段对的比较分值不会因片段的延伸而进一步升高。

为了分析上述分值随机性产生的几率大小，需要建立一个随机序列模型。对于蛋白质而言，最简单的序列模型可通过从一条序列中随机地选取氨基酸残基，当然这一条序列中各种残基的频率必需一定。另外，一对随机氨基酸的联配期望值必需为负值，否则不论联配片段是否相关的，都会得到高比值，统计理论也将派不上用场。

就象独立随机变量之和总是倾向于正态分布(normal distribution)一样，独立随机变量的最大值倾向于极值分布(extreme value distribution)。在研究最佳局部联配时，主要涉及的是后一种情况。在一定的序列长度 m 和 n 限定下，HSP 的统计值可由 2 个参数(k 和 λ)确定。最简单的形式，即不小于比较值为 S 的 HSP 个数，可由下列公式算得其期望值：

$$E = kmne^{-\lambda s} \qquad (3.7)$$

我们称该期望值为比值 S 的 E 值(E-Value)。

上述公式非常灵敏。在给定比值的情况下，将比较序列长度加倍，则 HSP 数(即 E 值)也将加倍，同样，S 值为 2X 的某个 HSP 长度必是 S 值为 X 的两倍，所以 E 值将随着 s 值的增大急剧减少。参数 K 和 λ 可分别被简单地视为搜索步长(search spacesize)和计分系统(scoring system)的特征数。

1.二进制值或标准比值(Bit score)

最初获得的比值(S)在没有计分系统或统计量 K 和 λ 的辅助下，没有什么意义。单独的比值就如同没有单位(米或者光年)的距离。可使比值按下式标准化：

$$S' = \frac{\lambda s - \ln k}{\ln 2} \qquad (3.8)$$

获得S’值就如同得到了具有标准单位的数值。

E 值因此可简化为：

$$E = mn2^{-S'} \qquad (3.9)$$

二进制值使所使用的计分系统赋予了统计学意义，使除了可以确定搜索步长外，同样可以计算相应的显著水平。

2.P 值(P-Value)(概率值)

具有大于或等于某一比值 S 的随机 HSP 数可由泊松分布(Poisson distribution)确定。由此可以计算出搜索到某一比值大于或等于 S 的 HSP 的机率为

---

[3]译自NCBI BLAST TURORIAL：The statistics of sequence similarity scores.

$$e^{-E}\frac{E^X}{X!} \qquad (3.10)$$

式中 E 由(3.7)式确定。

作为一个特例,搜索不到比值≥S的HSP概率为e^{-E},所以至少发现一个HSP (比值≥S)的概率为

$$P = 1 - e^{-E} = 1 - \exp(-kmne^{-\lambda x}) \qquad (3.11)$$

这是与比值 S 相关的 P 值(概率值)。例如,在可能搜索到 3 个比值≥S 的 HSP 的情况下，至少发现一个 HSP 的机率为 0.95[可由(3.11)式算得]。BLAST 程序中使用了 E 值而非 P 值，这主要是从直观和便于理解的角度考虑。比如 E 值等于 5 和 10，总比 P 值等于 0.993 和 0.99995 更直观。但是当 E<0.01 时，P 值与 E 值接近相同。

3.数据库搜索策略

E 值计算公式[公式(3.7)]可以应用于 2 个蛋白质序列长度分别为 m 和 n 的比较，但是对于某一序列长度为 m 的蛋白序列，如何在那些长短不一的数据库序列中找到与之匹配良好的序列呢？一种思路是把数据库中的所有蛋白序列与待查序列的关系都视为相同重要，也就是说对于 E 值均较低的短和长序列，它们是等同重要的。FASTA 程序近期版本便是采用这一策略。另一种思路是把长序列视为比短序列更重要，因为长序列往往包括更多的特异功能域(domain)。如果对序列长度上进行相关优先处理，则在计算数据库序列长度为 n 的 E 值时，将乘以 N/n，其中 N 为数据库中序列的总长度。根据公式(3.7)，E 值的计算可简单地把整个数据库序列视为长度为 N 的单条序列。BLAST 程序采用了这一策略。FASTA 策略中 E 值的计算还需再乘上数据库的序列条数。如果考虑到核酸数据库的序列长度变化更大,则在 DNA 序列相似性搜索时,BLAST 的策略可能会是合理的选择。

一些数据库搜索程序,例如 FASTA 或其它基于 Smith-Waterman 算法的程序,在进行序列搜索时，会对数据库中的每条序列进行联配并给出联配值，这些值大部分与未知序列无关，但它们被用于了 K 和 λ 参数的估计。这一方法避免了随机序列模型因使用真实序列(real sequence)造成的随意性，但同时产生了使用相关序列估计参数的难题。BLAST 仅通过部分而不是全部无关序列计算最适联配值，这赢得了搜索速度。因此，对于某一选定的替换矩阵和空位罚值，必须进行 K 和 λ 参数的预先估计，估计中使用真实序列，而非通过随机序列模型产生的模拟序列。这一估计的结果看来非常准确。

4.空位联配(gapped alignment)的统计问题

根据统计理论，以上述及的统计方法只适用于不含有空位的局部联配(非空位联配)。但是，许多计算试验和分析结果充分证明，上述统计方法同样适用于空位联配。对于非空位联配，可用基于替换矩阵和比较序列的残基频率的办法估计统计参数；对于空位联配，参数的估计则必须根据“ 随机 ”序列的大尺度比较。

5.边际效应(edge effect)

以上统计学方法对于短序列来说有些偏差。这些统计方法的基础理论是一个渐近理论，该理论假设局部联配可以适用于任何规模的联配。但是，一个高比值联配必须有一定的长度，不能从接近二条序列末端的地方开始。这种边际效应可以通过计算序列的“ 效应长度 ”(effective length)来修正。BLAST 程序中包含了这一修正过程。对于长于 200 残基的序列可以不进行边际效应的修正。

6.替换矩阵的选择

局部联配的结果与所选用的替换矩阵紧密相关。没有任何一个计分方案(即替换矩阵)可以适用于所有研究目标，对于局部联配的计分基础理论的正确理解可以极大促进序列分析准确性。相关内容详见第 4 小节。

7. 空位罚值(gap penalties)

联配中另一个重要问题是空位问题。空位处理是针对序列进化过程中可能发生的插入和缺失而设计的。插入和缺失可能只涉及 1 个或 2 个残基，也可能是整个功能域(domain)，所以，在进行空位罚值设计时必须反映这些情况。

有 2 个参数应用于空位罚值设定，一个与空位设置(gap opening)有关，另一个与空位扩展(gap extension)有关。任一空位的出现均处以空位设置罚值，而任一空位的扩大必须处于空位扩展罚值。对于一个空位长度为k的罚值W_k可用下式表示：

$$w_k = a + bk \tag{3.12}$$

其中 a 是空位设置罚值，b 为空位扩展罚值。这两个参数值设置的变化对联配产生影响(表 3.10).

表 3.10　空位设置和空位扩展罚值对联配的影响

空位设置罚值 (Gap opening penalty)	空位扩展罚值 (Gap extension penalty)	说　明 (Comment)
大	大	极少插入或缺失：适用于非常相关蛋白质间的联配；
大	小	少量大块插入：用于整个功能域可能插入的情况
小	大	大量小块插入：适用于亲缘关系较远的蛋白质同源性分析

经过多年的试验，一个合适的空位罚值已经被确定下来。大多数联配程序均对特定的替换矩阵设定了空位罚值的缺略值(default)，如果使用者希望使用不同的替换矩阵，则原来的空位罚值设定不一定合适。如何设定罚值并无明确的理论可遁，但大的空位设置罚值配以很小的空位扩展罚值被普遍证实是最佳的设定思路。

四．替换矩阵[4]

1.替换矩阵的一般原理

我们并不能直接计算出两条序列的最佳联配，我们需要找到一个可以估计任何联配的某一统计数，使生物学关系匹配最显著的联配统计数最大。

[4]本部分内容主要取自Weir B. S.（徐云碧等译）. 遗传学数据分析—群体遗传学离散型数据分析方法，北京：中国农业出版社，1996

先看以下 2 条氨基酸序列的联配情况。如果我们将各残基按相同的统计数处理，则 2 种联配(a 和 b)的得分将是相等的(9 个残基中 5 个匹配)：

```
(a)  TTYGAPPWCS              (b)  TTYGAPPWCS
      TGYAPPPWS                   TGYAPPPWS
      *  ***  *                   * *  ***
```

但是联配 a 是一些相对普通的残基(A、P、S 和 T)保持一致，而联配 b 则是一些相对稀有残基(W-色氨酸、Y-酪氨酸)相一致。我们需要一个更科学的赋分方法来反映匹配氨基酸间生物学和化学关系。

在联配中，C-C 匹配相对比 S-S 匹配更重要些，因为半胱氨酸(C)是具有非常特殊性质的相对稀有氨基酸，而丝氨酸(S)则相对普通。同样 D-E 匹配应取正值，因为这两个残基具有相同的化学性质，在两条联配的蛋白质序列中能起到相同的功用。但是，V-K 匹配则应被罚分，因为这两个残基毫无相似，不可能在两条序列中引到一样的作用。

替换矩阵(substitution matrices)包括了在联配中各种匹配方式如何赋分的信息，故替换矩阵又常被称为计分矩阵(scoring matrices)。

用于 DNA 序列联配的替换矩阵相对比较直观。以下是一个常被使用的替换矩阵：

	A	C	G	T
A	0.9	-0.1	-0.1	-0.1
C	-0.1	0.9	-0.1	-0.1
G	-0.1	-0.1	0.9	-0.1
T	-0.1	-0.1	-0.1	0.9

矩阵中每个匹配的碱基对均计为 0.9 分，每个不匹配的碱基对被罚 0.1 分，这样，下面一个联配的得分应为 4.3(=5×0.9+2×(-0.1))：

```
GCGCCTC
GCGGGTC
***  **
```

用于蛋白质联配的替换矩阵要复杂一些，因为没有一个矩阵可以适用各种情况。构建矩阵时应考虑不同的蛋白质家族在进化过程中，一种氨基酸突变成另一种氨基酸概率的差异，根据不同的蛋白质家族和预期的相似程度构建不同的替换矩阵。2 个最有名的蛋白质替换矩阵是 PAM 和 BLOSUM，它们分别是在 1979 年和 1992 年完成的。

最后，一个重要的概念必须明确。同源性(homology)和相似性(similarity)是不同的 2 个概念，不能混淆和混用。2 条序列具有同源性，意味着这两条序列有进化方面的关系，它们从一条共同的祖先序列进化而来；而相似性，只是表明一种相似程度。

2.PAM 氨基酸替换矩阵

在进行蛋白质序列联配时，必须通过一定的方法给联配的残基对赋予一定的分值，替换矩阵便是其中最重要的方法。

已故 Dayhoff 是蛋白质列序比较的先驱，她和她的同事们通过对蛋白质进化模式的研究，建立了一组被广泛应用的替换矩阵，这些矩阵常被称为 Dayhoff，MDM(Mutation Data Matrix)或 PAM(Percent Accepted Mutation)矩阵。

应用于DNA序列的许多算法最初是从氨基酸蛋白质序列的一些算法发展而来的。由于蛋白质最有可能是自然选择的目标，可以认为蛋白质序列的分析比DNA分析更具有生物学意义。蛋白质分析完全避免了几个三联体可能编码同一氨基酸的遗传密码简并问题。有必要进一步分析各种氨基酸间的同源性程度，以及在进化过程中一种氨基酸被另一种氨基酸替换的概率大小。也许把氨基酸按一定特性分成若干组更便于以上分析，例如氨基酸可分成中性疏水(G、A、V、L、I、F、P、M)、中性亲水(S、T、Y、W、N、E、C)、碱性(K、R、H)和酸性(D、E)氨基酸等。在比较许多具有相似性蛋白质序列的基础上，Dayhoff等于 1979 年构建了一个突变概率矩阵M(mutation probability matrix)。最初她们比较了许多对蛋白质序列，以确定进化过程中一种氨基酸被另一种氨基酸取代的经验资料。她们共观测到 1572 次取代“事件”。以此为基础，她们建立了表 3.11 的“可观测点突变矩阵”A(accepted point mutation matrix)(由于舍入误差使表中的数值相加不完全等于 1572)。氨基酸i被氨基酸j替换的经验次数(记作A_{ij})可从上表中找到。矩阵A可被称为原始PAM矩阵。

由矩阵A可以进一步获得突变概率矩阵M。矩阵M的元素M_{ij}表示经过一定的进化时期氨基酸j被氨基酸i所替换的经验频率。Dayhoff等进而把可观测突变百分率(percent accepted mutation或point accepted mutation per 100 residues)，即PAM作为一种时间度量单位。假设同一位点不会发生二次以上的突变，则 1PAM 等于 100 个氨基酸多肽链中预期发生一次替换所需的时间。

Dayhoff 提出了一个称为相对“突变力”(mutability)的概念，并将氨基酸 j 的相对突变力定义为观测到的氨基酸突变数除以联配序列中 j 氨基酸的频率，即：

$$m_j \propto \sum_{i \neq j} A_{ij} / f_j \qquad (3.13)$$

这里将氨基酸a_j所有可能的变化均考虑在内。各种氨基酸的m_j和f_j值(经标准化)列于表 3.12。

表 3.11 氨基酸替换次数表（Dayhof 等,1979）

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y
R	30																		
N	109	17																	
D	154	0	532																
C	33	10	0	0															
Q	93	120	50	76	0														
E	266	0	94	831	0	422													
G	579	10	156	162	10	30	112												
H	21	103	226	43	10	243	23	10											
I	66	30	36	13	17	8	35	0	3										
L	95	17	37	0	0	75	15	17	40	253									
K	57	477	322	85	0	147	104	60	23	43	39								
M	29	17	0	0	0	20	7	7	0	57	207	90							
F	20	7	7	0	0	0	0	17	20	90	167	0	17						
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7					
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269				
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696			
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0		
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6	
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17

注：总计观测到 1572 次替换；表中次数均已乘 10；祖先序列不明时，次数以平分处理

表 3.12 根据可观测点突变资料得到的氨基酸相对突变力(m_i)和频率f_i（Dayhoff等，1979）

	m_i	f_i		m_i	f_i
A	100	0.087	L	40	0.085
R	65	0.041	K	56	0.081
N	134	0.040	M	94	0.015
D	106	0.047	F	41	0.040
C	20	0.033	P	56	0.051
Q	93	0.038	S	120	0.070
E	102	0.050	T	97	0.058
G	49	0.089	W	18	0.010
H	66	0.034	Y	41	0.030
I	96	0.037	V	20	0.065

氨基酸a_j发生变化的概率为$1-M_{jj}$，这必须与突变力相一致，即

$$1-M_{jj} \propto m_j$$

或按下式定义常数λ：

$$M_{jj} = 1-\lambda m_j \qquad (3.14)$$

同样

$$M_{ij} \propto m_j A_{ij}$$

由于M_{jj}和$\sum_{k\neq j} M_{kj}$之和必为 1

$$M_{ij} = \lambda m_j A_{ij} / \sum_{k \neq j} A_{kj} \qquad (3.15)$$

又因 1PAM 为 100 氨基酸中预期发生一次替换，则另外 99 个氨基酸不发生变化，有

$$99 = 100 \sum_i f_i M_{ii}$$

$$\lambda = \frac{1}{100 \sum_i m_i f_i} \qquad (3.16)$$

Schwarts 和 Dayhoff(1979)发现将突变概率矩阵 M 250 次方处理获得的 250PAM 矩阵(表 3.13)，对于研究远缘蛋白质之间进化关系是一个合适的时间单位。

Dayhoff 等(1979)进一步定义了一个相对概率矩阵 R(relatedness odds matrix)，其元素 $R_{ij} = M_{ij} / f_i$。这一概率矩阵是对称的。该矩阵的元素已在类似 Needleman-Wunsch 算法中用作氨基酸 i 被氨基酸 j 替换的权重 w_{ij}，表 3.14 中各元素已经对数处理(故矩阵 R 又称为对数概率矩阵，Log-odds matrix)，并将最有可能发生相互替换的氨基酸归类排列。

表 3.13　250PAM 突变概率矩阵(Dayhoff 等,1979)

	A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
A	13	6	9	9	5	8	9	12	6	8	6	7	7	4	11	11	11	2	4	9
R	3	17	4	3	2	5	3	2	6	3	2	9	4	1	4	4	3	7	2	2
N	4	4	6	7	2	5	6	4	6	3	2	5	3	2	4	5	4	2	3	3
D	5	4	8	11	1	7	10	5	6	3	2	5	3	1	4	5	5	1	2	3
C	2	1	1	1	52	1	1	2	2	2	1	1	1	1	2	3	2	1	4	2
Q	3	5	5	6	1	10	7	3	7	2	3	5	3	1	4	3	3	1	2	3
E	5	4	7	11	1	9	12	5	6	3	2	5	3	1	4	5	5	1	2	3
G	12	5	10	10	4	7	9	27	5	5	4	6	5	3	8	11	9	2	3	7
H	2	5	5	4	2	7	4	2	15	2	2	3	2	2	3	3	2	2	3	2
I	3	2	2	2	2	2	2	2	2	10	6	2	6	5	2	3	4	1	3	9
L	6	4	4	3	2	6	4	3	5	15	34	4	20	13	5	4	6	6	7	13
K	6	18	10	8	2	10	8	5	8	5	4	24	9	2	6	8	8	4	3	5
M	1	1	1	1	0	1	1	1	1	2	3	2	6	2	1	1	1	1	1	2
F	2	1	2	1	1	1	1	1	3	5	6	1	4	32	1	2	2	4	20	3
P	7	5	5	4	3	5	4	5	5	3	3	4	3	2	20	6	5	1	2	4
S	9	6	8	7	7	6	7	9	6	5	4	7	5	3	9	10	9	4	4	6
T	8	5	6	6	4	5	5	6	4	6	4	6	5	3	6	8	11	2	3	6
W	0	2	0	0	0	0	0	0	1	0	1	0	0	1	0	1	0	55	1	0
Y	1	1	2	1	3	1	1	1	3	2	2	1	2	15	1	2	2	3	31	2
V	7	4	4	4	4	4	4	5	4	15	10	4	10	5	5	5	7	2	4	17

*表中数值均乘以 100；舍入误差使本表结果与上二表计算结果不完全相等。

表 3.14 250PAM 的对数概率矩阵(Dayhoff 等,1979)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W
C	12																			
S	0	2																		
T	-2	1	3																	
P	-3	1	0	6																
A	-2	1	1	1	2															
G	-3	1	0	-1	1	5														
N	-4	1	0	-1	0	0	2													
D	-5	0	0	-1	0	1	2	4												
E	-5	0	0	-1	0	0	1	3	4											
Q	-5	-1	-1	0	0	-1	1	2	2	4										
H	-3	-1	-1	0	-1	-2	2	1	1	3	6									
R	-4	0	-1	0	-2	-3	0	-1	-1	1	2	6								
K	-5	0	0	-1	-1	-2	1	0	0	1	0	3	5							
M	-5	-2	-1	-2	-1	-3	-2	-3	-2	-1	-2	0	0	6						
I	-2	-1	0	-2	-1	-3	-2	-2	-2	-2	-2	-2	-2	2	5					
L	-6	-3	-2	-3	-2	-4	-3	-4	-3	-2	-2	-3	-3	4	2	6				
V	-2	-1	0	-1	0	-1	-2	-2	-2	-2	-2	-2	-2	2	4	2	4			
F	-4	-3	-3	-5	-4	-5	-4	-6	-5	-5	-2	-4	-5	0	1	2	-1	9		
Y	0	-3	-3	-5	-3	-5	-2	-4	-4	-4	0	-4	-4	-2	-1	-1	-2	7	10	
W	-8	-2	-5	-6	-6	-7	-4	-7	-7	-5	-3	2	-3	-4	-5	-2	-6	0	0	17

*表中数值均乘以 10

1PAM 相当于所有的氨基酸平均有 1%发生了变化，经过 100PAM 的进化，并非每个氨基酸的残基均发生变化：有一些可能突变多次，甚至又变成原来的氨基酸，而另一些氨基酸可能根本没有发生过变化。这使我们认识到，利用大于 100PAM 的时间间隔可能达到区分同源性蛋白质的目的。应该注意，PAM 与进化时间之间没有大致对应关系，因为不同的蛋白质家族的进化速率是不同的。当 2 条序列进行相似性比较时，事先不知道怎样的进化时间(PAM)是恰当的。对于相近的序列，比较容易选择，即使不太合适的矩阵也无妨。在很多年里，PAM250(矩阵后面数字，如 PAM250、PAM100 等，表示一种进化的距离，数字越大，距离越远)是应用最广的替换矩阵，因为该矩阵是唯一由 Dayhoff 最初发表的矩阵。

后来一些学者利用大量新出现的蛋白质序列数据来更新 Dayhoff 最初计算的频率数值，由此新构建的 PAM 矩阵与最初的 PAM 矩阵没有太大的差异。

3.BLOSUM 氨基酸替换矩阵

另外一种构建矩阵的方法是由 Henikoff 等于 1992 年提出的，建成的矩阵为 BLOSUM(Blocks Substitution Matrices)。他们直接利用多序列联配(multiple alignment)分析亲缘关系较远的蛋白质，而不是用相近的序列。这方法的优点是符合实际观测结果，不足之处是它不能和进化挂起钩来。大量的试验表明，BLOSUM 矩阵总体比 PAM 矩阵更适合于生物学关系的分析和局部相似性搜索。

假设f_{ij}为序列联配中氨基酸i和j对(忽略顺序)，则i，j对氨基酸所占比例为：

$$q_{ij} = f_{ij} / \sum_{i,j} f_{ij} \qquad (3.17)$$

在完全独立的状况下，该比例的期望值为：

$$e_{ij}=\begin{cases}p_i^2 & i=j\\ 2p_ip_j & i\neq j\end{cases}$$

$$p_i=q_{ii}+\frac{1}{2}\sum_{j\neq i}q_{ij}$$

则 BLOSUM 矩阵元素(i,j)定义为：

$$s_{ij}=2\log_2(q_{ij}/e_{ij}) \qquad (3.18)$$

蛋白序列的高度保守区(highly conserved regions)或称为模块(block)数据被用于构建 BLOSUM 矩阵。BLOSUM 矩阵后的数字表示用于构建矩阵的模块的最小相似比例，例如 BLOSUM62 为用于构建矩阵的模块数据库中，序列片段的各联配点上至少 62%是相同的。矩阵后的数字越大，则表示关系越近。

4.DNA 替换矩阵

以上有关替换矩阵的讨论仅仅提及蛋白质序列的比较，但是，相关的原则同样适用于 DNA 序列的比较。在进行比较时应该意识到，用翻译而来的蛋白质序列总是好于直接用 DNA 序列。这是因为 DNA 序列的进化变化很少，在使用简单的 DNA 替换矩阵比较时，获得的同源性信息远少于蛋白质序列。

但是，有时我们希望比较一些非编码 DNA 序列。如前所述(见本小节第 1 部分)的 DNA 替换矩阵非常简单，所有 4 个碱基的匹配与不匹配的数值均设为相同，不同的只有匹配与否(0.9 和-0.1)。一个较复杂的模型是把转换(transition,两种嘧啶或两种嘌呤间的突变)频率设为高于颠换(tranversion,嘧啶与嘌呤间的突变)频率。

五．多序列联配

通过以上的两条序列算法，总是可以返回一个最佳匹配的联配结果。但是，当我们将两条以上的序列放在一起联配时，情况就就不一样了。现有实用的多序列联配方法还不能保证一定给出最优联配结果，只能给出一个近似值——往往人为的修正可以使联配结果更佳。人类(充满生物学智慧)的眼光在判断多序列联配方面远胜过目前的任何计算机。

同源序列的多序列联配是生物信息学一个重要课题。通过多序列联配结果，允许你观察残基可以改变到什么程度而蛋白质仍保持功能；它也可以使你得到围绕某一残基的三级结构信息。有关利用多序列联配预测蛋白结构的内容将在第六章讨论。有不少多序列联配程序可通过匿名 ftp 等服务获得，例如：ClustalW 等。

三条或三条以上序列的联配方法可分为几类，如用于两条序列联配的 Needleman-Munsch 等算法的改进算法、等级法(hierachical method)、片段法(segment method)、一致或区段法(consensus or regions'method)等。这些方法中，等级法是目前应用最为广泛的方法。

等级法又称为树法(tree method)，是由 Feng 和 Doolittge(1987)等人发展

的（ClustalW 程序）。由于两条序列的联配结果可以很容易地获得，多序列联配便可以在连续使用两条序列联配算法(如 Needleman-Wunsch 算法)基础上，通过先建“ 树 ”的思路来进行多序列联配。这一方法同样是一种动态规划方法。具体步骤如下：

①对所有序列进行两两联配分析，N 条序列应有 N×(N-1)/2 对；

②对两两联配的数据进行聚类分析，产生联配等级。该等级可用分叉树(binary tree)形式或简单的排序来表示；

③根据以上联配结果，首先从所有联配中相似性最好的两条序列开始，然后是剩余联配中相似性最好的两条序列……依次类推，直至多序列联配结束。一旦两条序列的联配被列入，则序列的位置就被固定下来。例如，对于序列 A、B、C、D，如果 A 与 C、B 与 D 分别是两两联配的最佳联配结果，则 A、B、C、D 四条序列的联配则通过比对 A-C 和 B-D 两个联配(每个联配位置取平均值)来确定。

这一组合方法对大量序列的多序列联配提供了实用的空位联配手段，除了最初的两序列间的联配过程，整个多序列联配过程是很快的。

可供同时联配多个序列的程序需要更多的计算机资源，而且不如前述的比联配序那么常用。在 GCG 的 PILEUP 程序中采用的 Feng 和 Doolittle 算法；NBRF 提供的 PIRAlign 是以 Needleman-Wunsch 算法的一个变通方案为基础。由 Greg Schuler 建立的 MACAW 程序则是适用于 Microsoft Windows(微软视窗)和 Macintosh 计算机的一个很有效的多序列联配软件。这些程序(见书后所附序列分析软件)都会从所有输入序列中找出共同区，然后以此为起点建立总体联配。如果将待比较的序列局限于它们的保守区域，这些程序一般较为有效。

应该指出，目前还没有一个最佳的多序列联配方法，自动联配程序给出的结果往往可以通过人为的分析而得到改进。

第三节 数据库搜索——BLAST 和 FASTA 应用

一．数据之海与一叶轻舟

《科学》(*science*)杂志在 2001 年 2 月 16 日的人类基因组专刊上发表了一篇题为“生物信息学：努力在数据的海洋里畅游”的文章，文章写到：“我们身处急速上涨的数据海洋中…我们如何避免没顶之灾？”一条可靠的办法可能是赶紧找到“一叶轻舟”，而且在轻舟上装上先进的电子设备，诸如卫星定位系统、卫星信息传输系统等等……BLAST 和 FASTA 便是这样的一条“轻舟”，它们往来穿梭，速度奇快。

比较和确定某一数据库中的序列与某一给定序列的相似性是生物信息学中最频繁使用和最有价值的操作。本质上这与两条序列的比较没有什么两样，只是要重复成千上万次。但是要严格地进行一次比较必定需要一定的耗时，所以必需考虑在一个合理的时间内完成搜索比较操作。目前有二个最为常用的程序服务于未知序列的数据库相似性搜索，即 BLAST 和 FASTA。FASTA 使用的是 Wilbur-Lipman 算法的改进算法，进行整体联配，重点查找那些可能达到匹配显著的联配。虽然 FASTA 不会错过那些匹配极好的序列，但有时会漏过一些匹配程度不高但达显著水平的序列。BLAST(Basic Local Alignment Search Tool，基本局部联配搜索工具)是基于匹配短序列片段，用一种强有力的统计模型来确定未知序列与数据库序列的最佳局部联配。

大多数研究目前都通过国际互联网 Internet 应用 NCBI 研制的 BLAST 程序(Basic Local Alignment Search Tool)来进行 DNA 和蛋白质序列相似性搜索。用一组 BLAST 程序联配可以快速进行核酸和蛋白质序列库的相似性检索。采用 BLAST 的基本算法编成了若干各不同的程序，分别使用特定的序列库和用于特定类型的输入序列。BLASTN 是在核苷酸序列库搜索核苷酸序列。BLASTP 是在蛋白质序列库中搜索氨基酸序列。TBLASTN 则可以在核酸序列库中搜索氨基酸序列，此时序列库在搜索之前要按所有 6 种读框即时翻译。与此相反的一项分析则由 BLASTX 来完成，它要将所输入的核酸序列按所有 6 种读框翻译，然后再以之搜索蛋白质序列库。近期 Altschul S.F.等人（1997）提出了一个通过寻找蛋白质家族保守序列来提高算法敏感性的 PSI-BLAST（Position-Specific Iterated BLAST）算法，并开发了相应的软件。PSI-BLAST 可以对数据库进行多轮循环检索，每一轮的检索速度都大约是 BLAST 的两倍，但每一轮都能提高检索的敏感性。它是目前 BLAST 程序家族中敏感性性最高的成员。

表 3.15 数据库相似性搜索程序 BLAST 和 FASTA 程序清单

程 序 (Program)	待检序列类型 (Probe type)	数据库类型	说 明 (Comment)
BLASTP	p	p	在蛋白质序列库中比对待检蛋白质序列
BLASTN	n	n	在核酸序列库中比对待核酸序列
BLASTX	n	p	在蛋白质序列库中比对待检核酸序列(用所有 6 种读框翻译)
TBLASTN	p	n	在核酸序列库(用 6 种读框即时翻译)中比对待检蛋白质序列
TBLASTX	n	n	在核酸序列库(用 6 种读框即时翻译)中比对待检核酸序列(同样用所有 6 种读框翻译)
FASTA3	p	p	在某一蛋白质序列库中搜索蛋白质相似序列
	n	n	在某一核酸序列库中搜索核酸相似序列
TFASTA3	p	n	在核酸序列库(已被即时翻译)中比对待检蛋白质序列
FASTX3	n	p	在蛋白质序列库中比对待检核酸序列(用 6 种读框翻译)
TFASTX3	p	n	在核酸序列库中比对待检蛋白质序列
SSEARCH	P/n	P/n	使用 Smith-Waterman 算法联配比对

注：n：核酸序列或核酸序列库；p：蛋白质序列或蛋白质序列库

如果目的序列中有蛋白质编码区，则用翻译的蛋白质序列来搜索蛋白质序列库要比用 DNA 序列搜索核酸序列库更有价值。由于蛋白质序列的进化要比 DNA 序列慢一些，在蛋白质序列水平上的远缘关系在 DNA 水平上可能被错过。如果无法确定编码区，则可利用 BLASTX 按所有 6 种读框来翻译 DNA 序列，然后用它搜索蛋白质序列库。由于蛋白质序列库仅包含已鉴定的蛋白质，所以必须采用 TBLASTN 程序在现有的 GenBank、EMBL 或 DDBJ DNA 序列库中检索新确定的氨基酸或翻译过的 DNA 序列。这种检索有时可以找到一些显著相似的 DNA 序列，而原本并不知道这些序列可编码蛋白质。

BLAST 的一项重要特性就是所报告的匹配序列的统计学显著性评分。这一统计学显著性评分是用 Karlin-Altschul 算法决定的，所算出的 Poisson 概率表明所得到的序列相似性随机出现的可能性。

另一个常用的核酸和蛋白质序列库搜索程序是 FASTA，即 FASTN 和 FASTP 程序的新版本。FASTA 首先在序列库中进行快速的初检，找出与待检序列高度相似的序列。这一快速检索局限于待检序列和序列库序列之间较短的完全相同序列区段上。

FASTA 首先要建立一个其长度由 K-tuple（ktup）值决定的所有可能的总表或字典。这一程序中使用的字长参数(或 K-tuple)表示所用的初始相配序列长度。

K-tuple 的大小可以变化并将间接影响搜索的速度和敏感度。然后程序要对待检序列和序列库中的所有序列进行处理，找出字典中长度与 K-trple 相等的所有序列段的位置。比较两个序列的字典要比比较两个序列本身快得多，可以有效地找出小段相似区。一旦通过初始的快速检索找到一批评分最高的序列，就可以仅对这些高分序列进行第二轮比较。第二轮的序列比对是采用 Needleman-Wunsch 算法(1970)进行空位联配计算，得出分析的最后结论。如果 FASTA 运行后找到较好的相似序列，有时采用较小的 K-tuple 值或换一个评分矩阵重新检索分析，也许会有帮助。

在终端计算机上 FASTA 检索的一个简便方法是使用电子邮件服务。有好几个机构都可以通过电子邮件自动地接受 FASTA 计算机检索要求，用电子邮件中所提供的序列，对多个序列库进行搜索，然后又通过电子邮件将结果送回。图 7.7.5 就是一人要求 FASTA 服务的电子邮件示范。BLAST 检索同样可以通过电子邮件或 Internet 服务进行。

```
TITLE A test search of the EMBL other Mammalian DNA sequences
LIB EMAM
WORD 4
LIST 100
ALIGN 20
SEQ
tgcttggctgaggagccataggacgagagcttcctggtgaaagtgtgtttcttgaaatcagcaccaccatg
gacagcaaa
END
```

图 3.6 送往EBI FASTA电子邮递服务中心（电子邮件地址：fasta@ebi.ac.uk）的一份邮件的内容。这份邮件要求用该序列对EMBL序列库中的其它哺乳动物序列进行检索。送回的答案中包括 100 条最匹配序列和头 20 条最匹配序列的联配结果。

不论是 FASTA 还是 TFASTA 都提供一项评分，以评价用前述 PAM250 矩阵生成的每一对联配序列的匹配程度。但 FASTA 并不像 BLAST 程序那样给出一项显著值。无论采用 FASTA 或 BLAST，推断相似性是否具有生物学意义都取决于研究者。要作出决断，必须充分考虑蛋白质已知的或推断的功能，与已知活性位点或模序的相似程度等等。

因为BLAST和FASTA采用不同的算法，同时用这两种搜索引擎重新检索某一特定序列往往是可取的。如果用其中一种找不到显著相似序列，不妨试一试另一程序。如果BLAST和FASTA均找不到显著匹配的序列，还可以选择第 3 条比较费时的搜索策略。一些网站允许用户使用基于Smith-Waterman算法的搜索程序，如 BLITZ。BLITZ(www.ebi.ac.uk/searchs/blitz.html)被设计在大型并行计算机上运行，因此使检索更灵敏。虽然运行这样的程序比较费时，但它们有时会发现一些被BLAST和FASTA错过的勉强达到显著的联配。

由于数据库相似性搜索是生物信息学最为重要的组成部分，所以很多网站都提供了 BLAST 和 FASTA 搜索服务。在选择何种数据源时，有很多标准可以应用。并非所有的 BLAST 和 FASTA 均提供相同的服务，你所搜索的数据库各不相同，这就如同我们有多种替换矩阵一样。另外，一些网站还为熟练使用者提供了特别服务。总之，在一些非冗余序列数据库中搜索均是被允许的。这类数据库至少包

括在 SWISS-PROT 和 PIR(蛋白质)或 EMBL 和 GenBank(核酸)的所有记录，这往往是最佳选择。但不要滥用这些资源，例如，如果你正在构建序列重叠群(contig)，则只需进行最终组合序列的 BLAST 或 FASTA 搜索即可，而不必对每个序列片段均进行搜索。同样，为了查找克隆载体的污染序列而进行整个非冗余数据库的 BLAST 运行，也不是一个有效办法。

二．BLAST：核苷酸数据库搜索

BLAST 包含有 5 个子程序，它是目前运行速度最快的检索搜索程序。最初的程序版本(Version1.4)不允许设置空位(gap)，这对运行速度的提高有好处。正如前文所述，空位直接关系到搜索结果，所以目前的 BLAST 版本(Version2.0)均能进行空位联配。BLAST 的快速得益于它的统计算法：BLAST 使用的是快速局部而不是缓慢、整体的联配策略。BLAST 不追求整条序列的匹配。

1．BLAST 实战操作(1)

如果这是你初次使用 BLAST，那不妨先按以下要求操作一次，先有个感性认识，然后再进一步了解和认识其细节：

在 Internet 中进入 EXPASY BLAST 主页

Basic BLAST | Advanced BLAST

Basic BLAST

Usage: Choose the the suitable BLAST program and database for your query sequence. Paste your sequence in one of the supported formats into the sequence field below and press the "Run BLAST" button. Don't forget your e-mail address, so that we can send you the results in case of traffic jam...

Make sure that the format button (next to the sequence field) shows the correct format .

See also our BLAST database description.

Please select the program:	blastp Program
Please select the database: ○ DNA databases	Please select
◉ Protein databases	Please select
☑ Gapped alignment on/off	blosum62 Select matrix
☑ BLAST filter on/off	Plain Text Select format
☑ Graphic output on/off	Query title (option)
Paste your sequence here: (or ID or accession number)	
required for tblast[nx] programs ->	E-mail address ☐ HTML

假如有一条人类基因序列，对这条序列我们一无所知。序列的提供者想对这

条序列进行常规分析来鉴定它。你可以这样进行：

①复制该序列：

AAAAGAAAAGGTTAGAAAGATGAGAGATGATAAAGGGTCCATTTGAGGTTAGGTAA
TATGGTTTGGTATCCCTGTAGTTAAAAGTTTTTGTCTTATTTTAGAATACTGTGAT
CTATTTCTTTAGTATTAATTTTTCCTTCTGTTTTCCTCATCTAGGGAACCCCAAGA
GCATCCAATAGAAGCTGTGCAATTATGTAAAATTTTCAACTGTCTTCCTCAAAATA
AAGAAGTATGGTAATCTTTACCTGTATACAGTGCAGAGCCTTCTCAGAAGCACAGA
ATATTTTTATATTTCCTTTATGTGAATTTTTAAGCTGCAAATCTGATGGCCTTAAT
TTCCTTTTTGACACTGAAAGTTTTGTAAAAGAAATCATGTCCATACACTTTGTTGC
AAGATGTGAATTATTGACACTGAACTTAATAACTGTGTACTGTTCGGAAGGGGTTC
CTCAAATTTTTTGACTTTTTTTGTATGTGTGTTTTTTCTTTTTTTTTAAGTTCTTA
TGAGGAGGGGAGGGTAAATAAACCACTGTGCGTCTTGGTGTAATTTGAAGATTGCC
CCATCTAGACTAGCAATCTCTTCATTATTCTCTGCTATATATAAAACGGTGCTGTG
AGGGAGGGGAAAAGCATTTTTCAATATATTGAACTTTTGTACTGAATTTTTTTGTA
ATAAGCAATCAAGGTTATAATTTTTTTTAAAATAGAAATTTTGTAAGAAGGCAATA
TTAACCTAATCACCATGTAAGCACTCTGGATGATGGATTCCACAAAACTTGGTTTT
ATGGTTACTTCTTCTCTTAGATTCTTAATTCATGAGGAGGGTGGGGGAGGGAGGTG
GAGGGAGGGAAGGGTTTCTCTATTAAAATGCATTCGTTGTGTTTTTTAAGATAGTG
TAACTTGCTTAAATTTCTTATGTGACATTAACAAATAAAAAAGCTCTTTTAATATTA
GATAA

②进入 EXPASY(EMBnet)BLAST 服务器主页。如果因故不能进入该服务器，也可使用其它网站的 BLAST 服务，但以下仅以 EXPASY BLAST 服务器为例；

③选择相关程序：BLASTN。该程序是在核酸数据库中进行相似核酸序列的搜索；

④选择数据库：EMBL without ESTs(DNA)。这是 EMBL 的主要核酸数据库；

⑤缺省替换矩阵选项：在 BLASTN 中不必应用矩阵；

⑥选择序列输入格式：Plain TEXT。以文本格式发送核酸序列；

⑦按如下选定： Gapped Alignment ON
BLAST filter: ON
Graphic Output: ON

⑧粘贴未知序列到输入框(Paste your sequence here：)内；

⑨按下运行按钮：Run BLAST；

⑩等待，并检查运行结果。

2．BLAST：结果报告

BLAST 的结果报告可能显得零乱，但是最主要的部分非常容易抓住。在报告的上部是有关程序的描述(如 BLASTN)、程序的版本和相关信息，接下来是你输入的未知序列，如果部分序列片段在过滤时未通过，则可看到一串 N 序列片段。再下来的几行提供了你搜索的数据库信息，包括该数据库最新更新时间。最后部分是在“searching”和“done”之间一系列(共 50 个)点(、)，如果是星号(*)，则表示程序在搜索该数据库时发生了障碍，少于 50 个点表明程序未能搜索整个数据库。这些因素必须予以考虑，你可能考虑重新运行一次。

在“Searching…Done”行下，你将看到一幅图。图最上面红色一条线代表未知的待搜索序列。在该线上有一个刻度，刻度下的数字为序列长度。其它不同颜色的线分别代表数据库中与之相似性显著的序列。可以看到，在本搜索进行的时候，数据库中只有一条与未知序列长度相仿的序列被列出，而其它找出的序列

均很短。由该图得出这样的结论：数据库中只有一条序列与你的未知序列有高度的相似性。

结果报告的再下面是一行一行的达到联配显著的序列描述。其中第一行(代表上图中与未知序列等长度的序列)如下：

emb|L37747|HSLAM11 [Homo sapiens]Homo sapiens lamin B1 gene, ex... 416 e-114

在这行描述中，E值(E-value)很重要，它是一行中最近面的一个数字(e-114)。E-114 可以表示为 1×10^{-114}或者说就是非常之趋近于零。这个数值表示你仅仅因为随机性造成获得这一联配结果的可能次数。这一数值越接近零，发生这一事件的可能性越小。从搜索的角度看，E值越小，联配结果越显著。

我们知道我们列举的序列来自人类，所以在以上结果中只有第一和第二行的序列是我们想要的。其它序列的 E 值较大，说明这些匹配结果很有可能是随机产生的，而且绝大部分序列来自其它生物。

注意 !“ Lamin β ”基因的序列很特别。当你搜索你自己的序列时，可能会得到 1 个以上匹配极好的序列，但是，统计上最显著的(E 值最小)并不总是你所要找的序列。应注意短的重复序列和模序家族(motif family)，它们可能不被统计联配算法(如 BLAST)看中。只把显著性当作一种导向，结合你的分子生物学知识和序列来源，你的人为判断能力在数据库搜索时总是有用的。

我们在报告中可以进一步看到实际的联配情况。它们的排序与上面各行的排序是一致的。一个短序列联配的例子：

```
Query:   1 ggccccaccacgccgctcag 20
             ||||||||||||||||| ||
Sbjct: 701 ggccccaccacgccgctgag 720
```

我们看到，未知序列与目标序列间几乎 100%匹配。一条竖线(|)连接两个碱基，表明它们是相同的。在未知序列中的第 18 个碱基 C 与数据库找到的匹配序列的第 18 个碱基 G 不相同，它们之间是空的，没有竖线。在其它的一些联配中，可以看到很大片的空缺。序列间不匹配除了缺少同源性外，还可能存在其它一些原因，如测序错误、未知序列的点突变等等。

我们这次的检索结果明白无误地告诉我们，第一条序列与我们的未知序列是一样的。回到报告中对序列的行描述部分，可以点击 EMBL 的身份号(HSLAM11)并查阅 EMBL 的数据库记录，记录中包括了该序列的相关信息，例如它所编码的蛋白质序列等。

3．BLAST 选项

我们回到 EXPASY BLAST 服务器主页，点击“ Advanced BLAST ”按钮，将出现一个有很多选项的页面。对于大多数搜索，最佳选项设置往往已被设为缺省状况，但是你可以方便地改变这些设置进行一些必要的搜索研究。我们需要准确地理解这些选项的真正含义。

“ WORDLENGTH ”(字长)选项：

BLAST 程序是通过比对未知序列与数据库序列中的短序列来发现最佳匹配序列的。最初进行“ 扫描 ”(scanning)就是确定匹配片段。序列的匹配程序由短

序列(定义为“word”,即字)的联配得分总和来决定。联配时,“字”的每个碱基均被计分：如果碱基对完全相同(如 A 与 A)，得某一正值；如果碱基对不很匹配(W 与 A 或 T)，则得某一略小的正值；如果两个碱基不匹配，则得一负值。总的合计得分便决定了序列间的相似程度。

得分高的匹配序列被称为高比值片段对(high-scoring segment pairs, HSP)。BLAST 程序在两个方向扩展 HSP，直至序列结束或联配已变为不显著。替换矩阵在扫描(scanning)和扩展过程被应用。最后在 BLAST 报告中被列出的序列都是所有得分最高的序列。

以上述及的初始字长便是由 W(WORDLENGTH)值设定。BLAST 只对字长为 W 的“字”进行扩展联配。BLAST 的字长缺省值为 11，即 BLASTN 将扫描数据库，直到发现那些与未知序列的 11 个连续碱基完全匹配的 11 个连续碱基长度片段为止。然后这些片段(即字)被扩展。11 个碱基的字长已能有效地排除中等分叉的同源性和几乎所有随机产生的显著联配。

“Filter”(过滤器)选项：

BLAST2.0 版本已有序列过滤器功能。过滤器将锁定诸如组成低复杂(low compositional complexity)序列区(如 Alu 序列)，用一系列 N(NNNNNN)替代这些程序。N 代表任意碱基(IUB-code)。只有未知待检序列被过滤替代，而数据库的序列将不被过滤。

过滤对绝大多数序列都是有益的,“Filter”项的缺省选项为 ON。例如，多 A 碱基的尾部和脯氨酸富积的序列，会得到人为的高联配得分而误导分析。这是因为这类序列数量极大，遍布整个基因组，直至整个数据库。

“Matrix”(矩阵)选项：

如前所述，联配的显著性是由返回的比对分值决定的，该分值反映的是所得到的联配随机产生的概率有多大。矩阵被用于鉴别数据库中的序列，同时又用来预测匹配的显著性大小。一般应接受运行程序推荐的矩阵。BLAST 系列程序主要使用两种类型矩阵(PAM 和 BLOSUM)。要准确地选择矩阵，必须了解矩阵和矩阵的具体计分方式。这方面的知识可参阅本章第二节“替换矩阵”部分。

注意！直接比较使用不同替换矩阵而获得的联配得分是没有意义的。同时，你可以为 BLAST、TBLASTN 或 TBLASTX 选择不同的矩阵，例如 PAM30、PAM70、BLOSUM80、BLOSUM62 等等，但是 BLASTN 不需要这些矩阵，搜索时，不必选定。

“EXPECT”选项：

你可能会想为搜索设定一个期望值阀值(EXPECT)，例如缺省值设为 10。这一设置则表示联配结果中将有 10 个匹配序列是由随机产生，如果联配的统计显著性值(E 值)小于该值(10)，则该联配将被检出，换句话说，比较低的阀值将使搜索的匹配要求更严格，结果报告中随机产生的匹配序列减少。

“Score Value”(分值)选项：

在“WORDLENGTH”选项中曾论及碱基对匹配程度的赋分问题，其赋分的标准可由分值选项的 M 和 N 两个参数设置。M 参数为匹配碱基的赋值，必需为一正整数；N 参数为不匹配碱基的赋值，必需为一负整数。

M/N 的比率决定了你所接受的进化分歧程度(degree of divergence)，M 和 N 的缺省值为 5 和-4。该比率(1.25)相当于在 100 个残基中约有 47 可以观测到的核酸点突变(PAM)。PAM 是被用来预测分子序列从祖先序列进化而来的程度。如果你调整 M 和 N 使比率提高,则 PAM 矩阵也应选择大些(指 PAM 矩阵后的数字)，以适应相应的较大的分歧程度。

输入框选项

你也许已注意到,在序列的输入框内可以键入 EMBL 的身份号(ID)或 GenBank 的记录号(accession number)。这样的输入选择将仅返回数据库中的某一序列资料(最新版本)，该序列与键入的记录号相对应。在不少情况下需要类似检索，例如核对 PCR 产物。其它一些选项情况可参阅 BLAST 的在线使用手册。

4．BLAST 实战操作(2)

写出以下几个问题的答案，然后与随后的答案比较一下：

①复制以下序列，运行 BLAST 程序搜索，鉴别该序列。除必需改变的设置，使用缺省设置。提示：你可能需要选择某一数据库和 BLAST 程序。

GTCCGGCCTGGGCGACAGAGCAAGACTCCGTCTCAAAAAAAAAAAAAAAAAAAAAAA
AAAAAAAAAAA

②该序列取自 GenBank 的一个记录(记录号为 S56967)。使用 BLAST 服务器找到该记录。

仔细察看以上序列，你会发现未能鉴别出该序列并没有什么奇怪。该序列是一条 Alu 序列！所以有如此多的匹配序列。

③再复制该序列并使用 BLASTX 程序。该程序是将待检序列翻译成蛋白质序列(6 种读框)，然后在蛋白质序列库 SWISS-PROT 中进行联配搜索。

选择 BLASTX、SWISS-PROT 等选项并运行后，检索结果与上次结果略有不同。分析得到的结果(BLAST2)，你可以知道该序列为 Alu 序列，你可能需要测定该基因的其它不同的片段或一条更长的扩展片段，以便能真正鉴别它。

④复制以下序列并运行 BLAST 搜索，检查检索结果。

GAATTCTAATCTCCCTCTCAACCCTACAGTCACCCATTTGGTATATTAAAGATGTGT
TGTCTACTGTCTAGTATCCCTCAAGTAGTGTCAGGAATTAGTCATTTAAATAGTCTG
CAAGCCAGGAGTGGTGGCTCATGTCT

你将能从检索结果中确定该序列编码的是人β-血球蛋白(beta heamoglobin)。在写作本书时，检索结果中前 2 条序列不仅匹配程度很好(100%和 99%同源)，而且它们与以上序列长度也一致。其它的匹配序列都很短。这很清楚地说明这是个β-血球蛋白基因，但是第二条序列中有一个 C 碱基因与第一条不相同，这提醒你在最后确定前应该再检索一下你的测序结果是否正确。

⑤问题④中的序列非常特异，如果同该序列的前 15 个碱基去搜索是什么样的结果？

GAATTCTAATCTCCCTCTCAACCC

没有发现任何线索！这很奇怪，因为我们刚刚进行了检索。这一情况告诉我们这样一个事实：检索结果中没有匹配的序列(“NO Hits”)并不一定是数据库中没有这些序列，而是可能因为检索标准设置的问题。

再复制这 15 个碱基序列，回到 BLAST 主页。点击“Advanced BLAST”(高级 BLAST)按钮，使用 EMBL 数据库“nr”亚类，E 值调整到 100，关掉(off)“XBLAST－repsim filter”过滤器，然后运行。

在写作本书时，结果中只有 2 个匹配序列，即以上搜索到的 2 条序列。由此可以确定该序列极有可能是β-血球蛋白基因。

三．BLAST：蛋白质数据库搜索

蛋白质数据库搜索是应该掌握的最重要的生物信息学技能，因为该搜索的灵敏性大约是核酸数据库搜索的 2-5 倍。蛋白质数据库搜索灵敏性好的原因包括：DNA 密码只有 4 个，在每个位置上的密码只有 4 种可能，而蛋白质有 20 种可能；遗传编码的多样性，n 个三联体密码编码一种氨基酸；虽然某一蛋白质序列与你的未知序列相同，但是你不能得到一个明确匹配的 DNA 序列。另外，蛋白质序列的相似性比 DNA 序列更保守。

蛋白质的直系同源性(orthologue)检索已越来越成为分子生物学的重要组成部分。目前一种酵母(*Sacharomyces cerevisiae*)和一种线虫(*Caenorhabditis elegans*)等的基因组序列已完成，同源性分析已在有效地进行中。如果人类的某一特异蛋白与以上的某一同源族相匹配，则可以确定该蛋白的可能功能，这将节省大量研究时间和经费。这一方面研究已有很好的例子(可参阅 Chervitz SA,et al. Comparison of complete protein sets of worm and yeast:orthology and divergence. *Science*.1998,282:2022-2028)。

两个主要蛋白质数据库(PIR 和 SWISS-PROT)的记录没有象三个主要核酸数据库一样相互交换。两个数据库各有优缺点，你必须考虑选择合适的数据库进行搜索。

下面我们进行 SWISS-PROT 数据库的进行实战搜索操作：

在以下的操作中，你将结合以上有关 BLAST 的知识，学会如何从相关数据库中获取信息。

①选择序列联配程序：你可能选择 BLAST 或 FASTA 服务器，本例选 BLAST。
②复制以下人类蛋白质序列：

MSTAVLENPGLGRKLSDFGQETSYIEDNCNQNGAISLIFSLKEEVGALAKVLR
LFEENDVNLTHIESRPSRLKKDEYEFFTHLDKRSLPALTNIIKILRHDIGATVHE
LSRDKKKDTVPWFPRTIQELDRFANQILSYGAELDADHPGFKDPVYRARRK
QFADIAYNYRHGQPIPRVEYMEEEKKTWGTVFKTLKSLYKTHACYEYNHIFP
LLEKYCGFHEDNIPQLEDVSQFLQTCTGFRLRPVAGLLSSRDFLGGLAFRVF
HCTQYIRHGSKPMYTPEPDICHELLGHVPLFSDRSFAQFSQEIGLASLGAPD
EYIEKLATIYWFTVEFGLCKQGDSIKAYGAGLLSSFGELQYCLSEKPKLLPLEL
EKTAIQNYTVTEFQPLYYVAESFNDAKEKVRNFAATIPRPFSVRYDPYTQRIE
VLDNTQQLKILADSINSEIGILCSALQKIK

③粘贴以上序列到输入框内并调整相关选项。你不必进行高级 BLAST 检索，但你必须选择数据库和程序。本例选 SWISS-PROT 数据库。
④运行 BLAST。使用提供的链接功能阅读检索结果报告。获得的报告可以用 NICE-PROT 阅读，其界面更友好和完整。
回答以下问题。你也许需要点击与 SWISS-PROT 报告链接的相关数据库信息。回答这些问题需要一定的时间，但它可以使你明白你能得到哪些信息并如何

得到它们。

问　　题	答　　案
该记录的 SWISS-PROT 名称是什么？	PH4H_Human
SWISS-PROT 最初的记录号是多少？	P00439
该蛋白的最普通用名？	Phenylalanine-4-Hydroxylase
该基因名称？	PAH
哪一年该催化功能区的晶体结构被确定？作者是谁？	1997,Erlandsen
该酶发挥功能是否需要协因子(co-factor)?如果是，是哪个因子？	是， ferrous ion
与该酶缺失直接有关的最普通疾病是什么？	Phenylketonuria(PKU)
该基因的细胞遗传学位点？(例如 13p10.1)	12q24.1
PAHdb 是什么？	PAH 突变体数据库
该蛋白质有多少氨基酸残基？	452
该蛋白质的分子重量是多少？	51.862kDa
如何得到该蛋白质的三维图象？	进入 PDB 数据库

获得的以上答案的正确操作：选择 BLASTP 和 SWISS-PROT，结果显示只有人类的 PAH1 序列与未知序列完全(100%)相同，由此可以确定未知序列。点击 PAH1 记录号，进入 SWISS-PROT 数据库，查阅该记录信息。在此处可用 NICE-PROT(点击)阅读。观察三维图象时，可在“Cross-references”(交叉文献)下点击 PDB 数据库链接按钮。

四．FASTA：另一种搜索策略

1．FASTA算法[1]

FASTA 的原型是 David Lipman 和 William Pearson(1985)提出的用于蛋白质同源比较的 FASTP。FASTA 提高了 FASTP 的灵感性但速度并没有损失多少(pearson and Lipman,1988)。它可以用来进行 DNA 对 DNA，DNA 对蛋白质(将 DNA 按 6 个读框“翻译”成氨基酸序列，再与蛋白质比较)和蛋白质对蛋白质的同源比较。下面以两条氨基酸序列的比较为例介绍算法的基本思路。

算法可以分为 4 步：

第一步：

FASTA 首先找出进行比较的两条序列所有长度为 K-tuple 的连续的一致序列片段。例如以下两条蛋白质序列：

序　列	位置 1	2	3	4	5	6	7
1	F	L	W	R	T	W	S
2	T	W	K	T	W	T	

设K-tuple =2，则序列 2 中有两个符合条件的片段(用下划线表示)，相对于序列 1 的偏移(offset)分别是 4 和 1 [对于一对开始位置为(x_1,x_2)的一致片段，偏移

[1]本部分内容主要取自F．奥斯伯，R．E．金斯顿等．精编分子生物学实验指南，北京：科学出版社，1998

定义为x_1-x_2。在上例中有两对(x_1,x_2)，即（5,1）和(5,4)]。这种片段的一致性可以表示为对角线图,两条序列中的一对一致片段在图中表示为一段对角线。(图3.5)。

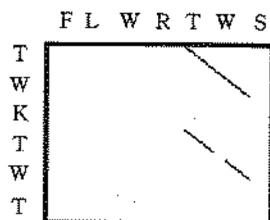


图 3.5 序列 FLWRTW 和 STWKTWT 比较形成的对角线图

本例是两条非常短的氨基酸序列，在实际比较长的蛋白质序列或 DNA 序列时，对角线图如图 3.6A 所示。

对于图中每一条完整的对角线(即同一偏移)上的一致片段，如果片段间距小于用户界定的界限，则将片段连接起来作为一条一致片段。对这些片段进行计分，每一对对应的元素，一致的加分，不一致的扣分。完成了所有一致片段的计分后，选出 10 条分值最高的片段进入下一轮计算，如图 3.6B.

第 2 步：

FASTA 将这 10 对片段重新计分。这轮计分允许保守突变，对蛋白质来就，就是使用 PAM250 等替换矩阵。简单地说，替换矩阵就是对应于 20×20 种氨基酸替换(比如 R 替换成 P)的计分规则所构成的 20×20 的矩阵。这种矩阵是从蛋白质进化实例中总结出来的经验矩阵,它给予进化上相对保守的氨基酸替换比非保守的替换更高的分值。在重新计算分值后，在每一条这样的片段中找出分值最高的子片段，作为“初始区域”(initial region)进入下一步。在 initial region 中，最高的分值计为 initl。

第 3 步：

在这一步中，FASTA 选出分值高于用户确定的界限且相互之间不重叠的初始区域，并尝试将这些初始区域连接起来。当然，由于连接而出现的缺失和插入情况要作相应的扣分。FASTA 在这一步才考虑插入和缺失的情况，最终找出能够得到的最高分值的初始区域或连接起来的数个初始区域。这一步计算出的最高分计为 initn。见图 3.6C。

第 4 步：

以 initl 片段或(initn 的片段)为中心，向前后延伸一定的长度。在这样一个区域中(见图3.6D中虚线间的区域),应用Smith-Waterman算法进行重新对齐，最终的得分计为 opt。

在实际操作中,用户可以在需要达到的灵感性程度和所需时间之间进行权衡(一般来说，要达到更高的敏感性总是需要更长的运算时间)，决定采用 initn 还是 opt 作为两条序列相似程度的分值。研究表明:使用 initn 与使用 opt 相比，前者损失的敏感性并不太大，但运算速度却快得多(Pearson WR,1991)。

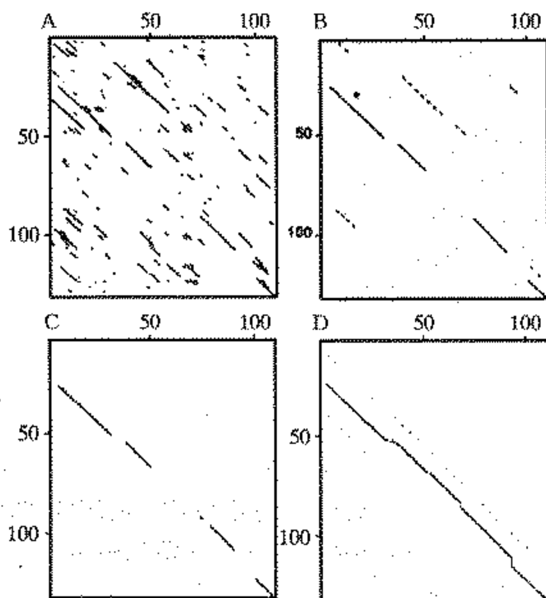


图 3.6 FASTA 的 4 步（A→D）算法图示

2.FASTA 选项

只有 10 个碱基长度的短序列可以用 FASTA 检索。其检索速度主要由 KTUP 值决定，该值用于限定字长。在 BLAST 中，“字”(word)表示用于联配的短序列片段，高比值的字(HSP)被选定并进行扩展。在 FASTA 中，字不被计分，联配只有在完全匹配的情况下再继续下去。

“Matrix”(矩阵)选项：

FASTA 和 BLAST 在最初扫描和扩展(FASTA 只使用扩展)阶段主要的不同之处是 FAST 允许多义密码子(IUB ambiguity codes)包括其中。大多数 FASTA 服务器提供了 BLOSUM 或 PAM 系列替换矩阵。FASTA 的缺省推荐矩阵是 BLOSUM50。如果 BLOSUM50 不合适，BLOSUM62 是另一个可行的选择。如果你在进行突变性质的进化分析时，不妨试试 PAM。

Smith-Waterman 算法：

你可以选择比以上讨论的算法更严格的算法进行联配。SSEARCH3 程序使用了 Smith-Waterman 算法，适用于高精细的检索，但运行速度非常慢，但是在你有类似搜索需要时，它无疑是值得应用的。往往要求键入一个 E-mail 地址，以便将搜索结果通过 E-mail 发送给你。

“KTUP”选项：

KTUP 值(字长)可在 1-6 整数之间选择。KTUP=2 的选项设置将是 KTUP=1 的设置搜索速度快 5 倍，因为 KUTP=1 时，服务器将对每个碱基进行联配，而 KUTP=2 时，则以 2 个碱基进行联配。有些服务器限制 KUTP 设置必须在 3-6 之间。缺省设置为 KUTP=6。

“GAPOPEN”和“GAPTEXT”(空位设置与空位扩展)选项：

与 BLAST 一样，空位设置和空位扩展的罚值必须为负值，它们的缺省设置分别为-16 和-4。注意：FASTX 和 TFSTX 对移码(frameshift)也可以设置罚值。

“STRAND”(转向)选项：

正常情况下，STRAND 设置为“upper”，另外选项为“bottom”。如果选

“ bottom ”，则 FASTA 将对未知待检序列转向后再进行搜索。

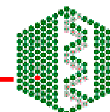
3.FASTA 的实战操作及其结果报告

如果你已经进行过 BLAST 的实战操作并读懂了其结果报告，则 FASTA 就不在话下了。此处仅给出了一条例举序列及其 FASTA 结果报告，不妨自己试试。

进入EMBL的EBI网站FASTA3服务器，并复制下列未知DNA序列：

CCAGATCCTGGACAGAGGACAATGGCTTCCATGCAATTGGGCAGATGTGTGAGGCACCTGTGGTGACC

EMBL
European Bioinformatics Institute



Fasta3 Help Tools EBI Home Run Fasta3 RESET FORM

YOUR EMAIL	SEARCH TITLE	RESULTS	DNA STRAND	MATRIX
	Sequence	interactive	none	BLOSUM50
GAP PENALTIES	SCORES & ALIGNMENTS	KTUP/HISTOGRAM	PROGRAM	DATABASES
OPEN -12 RESIDUE -2	SCORES 50 ALIGNMENTS 50	KTUP 2 HIST no	fasta3 fastx3 fasty3 fastf3 fasts3	Protein swall swiss-prot swiss-new sptrembl

Enter or Paste a DNA/RNA Sequence in any format:

CCAGATCCTGGACAGAGGACAATGGCTTCCATGCAATTGGGCAGATGTGTGAGGCACCTG
TGGTGACC

首先选择 fasta3 程序、EHUM 数据库和 bottom(STRAND) 选项，运行 fasta3。可以得到如下结果：

```
FASTA (3.39 May 2001) function [optimized, +5/-4 matrix (5:-4)] ktup: 6
 join: 45, opt: 30, gap-pen: -16/ -4, width:  16
 Scan time: 54.270
The best scores are:                                       opt bits E(145773)
EM_HUM:AF015262 AF015262 Homo sapiens Down Syn (79920) [r]  125   36    0.83
EM_HUM:HS229043 AJ229043 Homo sapiens 959 kb c (48446) [r]  125   36    0.96

>>EM_HUM:AF015262 AF015262 Homo sapiens Down Syndrome cr  (79920 nt)
rev-comp initn:  74 init1:  74 opt: 125  Z-score: 120.7  bits: 36.3 E(): 0.83
 67.164% identity (68.182% ungapped) in 67 nt overlap (68-2:209228-209293)

                                        60        50        40
EMBOS-                               GGTCACCACAGGTGCCTCACACATCTGCCC
                                     :::: ::: :  : :: ::: :: ::::::
EM_HUM GCACCAACCGTGTTCCAGGCTCTCTCAGGTGGTCTCCATAACTACCCCACTCACCTGCCC
           209200    209210    209220    209230    209240    209250

                 30        20        10
EMBOS- AATTGCATGGAAGCCATTGTCCTCTGTCCAGGATCTGG
       :::  : : :::::::   :  : : : : :: ::::
```

有两个基因序列（AF015262 和 HS229043）报告，但它们与未知序列的碱基相同率均不到 70%。

重新进行 fasta 选项设置：选 fastx 程序和 SWISS-PROT 数据库，并重新运

行 fasta3 ，得到以下结果：

```
FASTX (3.39 May 2001) function [optimized, BL50 matrix (15:-5:-1)] ktup: 2
 join: 36, opt: 30, gap-pen: -12/ -2 shift: -20, width:  16
 Scan time:  2.150
The best scores are:                                         opt bits E(96664)
SW:BRC1_HUMAN P38398 BREAST CANCER TYPE 1 SUSC (1863) [f]  149    55 2.6e-07
SW:BRC1_CANFA Q95153 BREAST CANCER TYPE 1 SUSC (1878) [f]  143    53   1e-06
SW:NODL_RHIME P28266 NODULATION PROTEIN L (EC  ( 183) [f]   70    29     2.2

>>SW:BRC1_HUMAN P38398 BREAST CANCER TYPE 1 SUSCEPTIBILI  (1863 aa)
 initn: 148 init1: 148 opt: 149  Z-score: 253.3  bits: 55.4 E(): 2.6e-07
Smith-Waterman score: 149;  95.238% identity (95.238% ungapped) in 21 aa overlap

              30        60
EMBOSS SWTEDNGFHAIGQMCEAPVVT
       .::::::::::::::::::::
SW:BRC AWTEDNGFHAIGQMCEAPVVT
          1820      1830
```

得到两个匹配良好的基因序列（P38398 和 Q95153），碱基相同率均在 90%以上。应如何确定未知序列和解释以上两次搜索结果？

第四节　　寡核苷酸设计

有关序列分析的内容非常丰富，本节只对引物设计进行简单讨论，而其它一些内容(如 ORF 的查找)在下章中论述。

一．寡核苷酸设计

聚合酶链式反应(PCR)技术的广泛应用，刺激了多种辅助设计和用于PCR的寡核苷酸引物程序的兴起。一些程序可通过Internet免费索取，例如Primer、OSP、PGEN、Amplify等(见附录)。一般而言，这些程序通过检索已知的重复序列元件，然后再分析假定引物的长度和GC含量从而优化T_m值，实现PCR引物的辅助设计。

1.引物设计

许多软件可以根据相应的标准为你的序列设计引物。如果你熟悉 PCR，你将理解软件中的有关选项；如果不熟悉，相关软件中均会有使用手册备查。以下是应用 Primer3 程序(见 EMBnet 挪威站点)进行的一次引物设计，注意设计结果中一些有用的信息(如 G-C 组成比率、建议的退火温度等)：

例举序列：

GACTGTGGCTGCTGGCGTTGAGGGAAACCTGCCTGTACGTGAGGCCCTAAAAAGCCA
GAGACCTCACTCCCGGGGAGCCAGCATGTCCACTGCGGTCCTGGAAAACCCAGGCTT
GGGCAGGAAACTCTCTGACTTTGGACAGGAAACAAGCTATATTGAAGACAACTGCAA
TCAAAATGGTGCCATATCACTGATCTTCTCACTCAAAGAAGAAGTTGGTGCATTGGC
CAAAGTATTGCGCTTATTTGAGGAGAATGATGTAAACCTGACCCACATTGAATCTAG
ACCTTCTCGTTTAAAGAAAGATGAGTATGAATTTTTCACCCATTTGGATAAACGTAG
CCTGCCTGCTCTGACAAACATCATCAAGATCTTGAGGCATGACATTGGTGCCACTGT
CCATGAGCTTTCACGAGATAA

结果：

```
No mispriming library specified
Using 1-based sequence positions
OLIGO            start  len     tm     gc%   any    3' seq
LEFT PRIMER        112   20  59.98   55.00  3.00  3.00 CTTGGGCAGGAAACTCTCTG
RIGHT PRIMER       364   20  59.99   50.00  3.00  3.00 GATGTTTGTCAGAGCAGGCA
SEQUENCE SIZE: 420
INCLUDED REGION SIZE: 420

PRODUCT SIZE: 253, PAIR ANY COMPL: 5.00, PAIR 3' COMPL: 2.00

    1 GACTGTGGCTGCTGGCGTTGAGGGAAACCTGCCTGTACGTGAGGCCCTAAAAAGCCAGAG


   61 ACCTCACTCCCGGGGAGCCAGCATGTCCACTGCGGTCCTGGAAAACCCAGGCTTGGGCAG
                                                         >>>>>>>>>

  121 GAAACTCTCTGACTTTGGACAGGAAACAAGCTATATTGAAGACAACTGCAATCAAAATGG
      >>>>>>>>>>>

  181 TGCCATATCACTGATCTTCTCACTCAAAGAAGAAGTTGGTGCATTGGCCAAAGTATTGCG


  241 CTTATTTGAGGAGAATGATGTAAACCTGACCCACATTGAATCTAGACCTTCTCGTTTAAA


  301 GAAAGATGAGTATGAATTTTTCACCCATTTGGATAAACGTAGCCTGCCTGCTCTGACAAA
                                                  <<<<<<<<<<<<<<<<

  361 CATCATCAAGATCTTGAGGCATGACATTGGTGCCACTGTCCATGAGCTTTCACGAGATAA
      <<<<

KEYS (in order of precedence):
>>>>>> left primer
<<<<<< right primer
```

用于测序或 PCR 的引物，需要选定可特异识别靶区的适当序列，然后检查该序列，以杜绝寡核苷酸形成稳定二级结构的可能。序列中的反向重复查找可通过找寻重复序列或 RNA 折叠的程序来进行。如果查出可能的茎区结构，引物序列可以向前或向后移动几个核苷酸，以期尽量削弱所预测形成的二级结构。寡核苷酸序列还应与适当的载体及插入 DNA 两条链上的序列进行比较。显而易见，测序引物应仅与靶 DNA 的一个区段相配对。若引物与靶 DNA 序列的非目标区很相似，即使只有一个位置不完全配对，这种情况一般也要避免。对于用于扩增基因组 DNA 的 PCR 引物，引物序列应与 GenBank 序列库中的序列进行比较，以检查是否有显著相似的配对区，如果寡核苷酸序列出现在任何已知 DNA 序列中，或者有更严重的情况，也就是寡核苷酸序列出现在任何已知的重复序列元件中，那么引物序列就必须改变。

2.用于检测相关基因的简并探针

一旦找出保守的蛋白质序列，并可以设计简并寡核苷酸作为杂交探针来筛选文库，找出蛋白质家族中的其他成员。设计这一用途的寡核苷酸，必须先将保守的蛋白质序列翻译成简并 DNA 序列。大多数软件包都可提供这一功能，其输出结果是采用 IUPCC 简并核苷酸代码表示的 DNA 序列。随后就可以合成对应于这一翻译蛋白质序列的简并寡核苷酸。多数 DNA 合成仪都可以合成除了末端核苷酸以外的序列内部任何位置上带有一种以上核苷酸的寡核苷酸。

第四章 基因组测序及分析

人类基因组和其它一些生物基因组的大规模测序将成为科学史上的一个里程碑。基因组测序带动了一大批相关学科和技术的发展，一批新兴学科脱颖而出，生物信息学、基因组学、蛋白质组学等便是一批最前沿的新兴学科。可以说，基因组测序及其序列分析使整个生命科学界的真正认识了生物信息学，生物信息学也真正成为了一门受到广泛重视的独立学科。

基因组测序及其分析实际是人类的又一场“淘金”和“探险”运动。哥伦布等一大批探险家在几百年前发现了美洲、澳洲等一大批新大陆，最终使人类认识了地球上的每一块处女地。于是有人形象地把人类目前的基因组研究形象地比喻为“地球探险”，并把基因组研究称为基因组地理（genomic geography）。我们不妨想象一下，人类基因组的各条染色体就如同人类基因“地球”上的 7 大洲，寻找新基因和搞清楚基因组结构与功能的过程恰如开垦地球上的每一块处女地，而这些处女地上可能蕴藏着无穷的宝藏。目前人类全基因组序列已基本测定完成，另有一大批生物也已完成基因组测定或正在进行。世界上无数大型测序仪（最好的测序仪一次可以阅读 1000 多个碱基）日夜不停地运转，每日获得的序列数据以百万和千万计。同时，来自政府和企业的大量投资，使整个世界的测序能力与日俱增。面对基因组的天文数据，分析方法举足轻重，大量新的分析方法被提出和改进，大量重要基因被发现；大量来自基因组水平上的分析比较结果被公布，这些结果正在改变人类已有的一些观念。

第一节 DNA 测序及序列片段的拼接

一．DNA测序的一般方法[1]

1．DNA 测序的基本原理

DNA 序列测定的工作基础是在变性聚丙烯酰胺凝胶(测序胶)上进行的高分离度的电泳过程。这些所谓的测序胶能在长达 500bp 的单链寡核苷酸中分辨出一个脱氧核苷酸的差异。操作时，在相应的待测 DNA 区段产生一套标记的寡核苷酸单链，它们有固定的起点，但另一端是按模板序列连续终止于各不相同的核苷酸。确定每个脱氧核糖核苷酸的序列的关键，是在 4 个独立的酶学或化学反应中产生终止于所有不同的 A、T、G、C 位点的寡核苷酸链，而这 4 个反应的寡核苷酸产物在测序胶的相邻泳道中都能被一一分辨出来。由于在 4 个泳道中再现了所有的可能寡核苷酸链，DNA 的序列能从图 4.1 所示的 4 个寡核苷酸“阶梯”中依次直接读出。

实际上，从一套测序反应中所能获得的信息量受限于测序胶的分离度。虽然最新的测序技术经常可从一套测序反应中测到高达 500 核苷酸的信息，但获得的可靠序列信息大约在 300 个核苷酸。因此，如果待测 DNA 的区段在 300 核苷酸以

[1]本部分内容主要取自F．奥斯伯，R．E．金斯顿等．精编分子生物学实验指南，北京：科学出版社，1998

内，所需的工作只是简单地将此片段克隆于合适的载体，以产生一个能方便地进行测序的重组 DNA 分子。

对于大片段 DNA 的序列测定，往往需要将其切割成能单独进行测定的小片段，这可通过随机的或有序的方式进行。下一节将讨论测定大片段 DNA 的策略。

目前广泛应用于 DNA 序列测定的方法有酶学的双脱氧法和化学裂解法，在产生寡核苷酸“ 阶梯 ”的技术上，两者截然不同。酶学双脱氧法是利用 DNA 聚合酶合成与模板互补的标记拷贝，化学裂解法是一套碱基专一的化学试剂作用于标记好的 DNA 链。这两种方法下面将进一步描述。

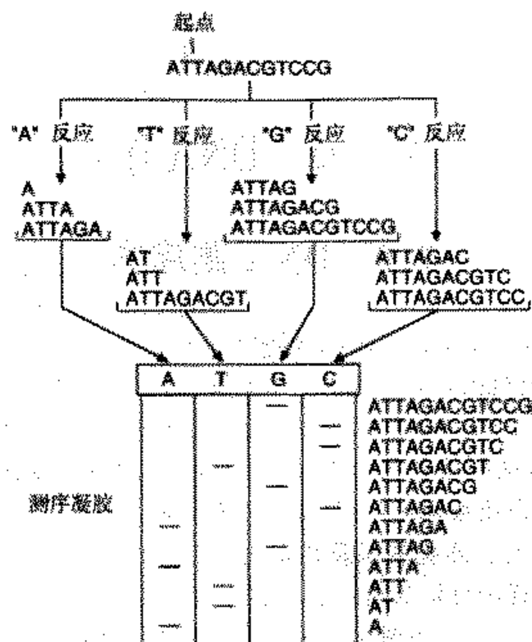


图 4.1 DNA 测序的一般策略。进行 DNA 序列测定时，在 4 个独立的反应中，各产生一套放射性标记的单链寡核苷酸，它们有固定的起点，另一端终止于不同的 A、T、G 或 C 位点。每个反应的产物在高分离度的聚丙烯酰胺凝胶上电泳分级。经放射自显影，DNA 序列可从凝胶上直接读出（奥斯伯等，1998）。

2．双脱氧测定法(Sanger 法)

双脱氧法或酶法利用DNA聚合酶合成单链DNA模板的互补拷贝，这一方法最先（1977）由F．Sanger 及其合作者提出。DNA聚合酶不能起始DNA链的合成，而能在退火于“ 模板 ”DNA的引物 3ˊ端上进行链的延伸（如图 4.2）。通过与模板碱基的特异性配对，脱氧核糖核苷酸(dNTP)被掺入到引物的生长链上。链的延伸是通过引物生长端的 3ˊ羟基与被掺入脱氧核糖核苷酸的 5ˊ磷酸基的反应形成磷酸二酯键，在总体上看，链是从 5ˊ→3ˊ方向延伸的。

双脱氧测序法利用了DNA聚合酶能从双脱氧核糖核苷酸(ddNTP)为底物的特性。当ddNTP被掺入到延伸着的引物的 3ˊ端时，由于链上 3ˊ羟基的缺如，链的延伸就终止于G、A、T或C。在 4 个测序反应中，每个反应只需各加入 4 种可能的ddNTP中的一种，就将产生如图 4.1 所示的 4 个序列阶梯。调整每个测反应中的ddNTP与dNTP的比例，使引物的延伸在对应于模板DNA上的每个可能掺入ddNTP的位置都

有可能发生终止。以这种测序方式，每个延伸反应的产物是一系列长短不一的引物延伸链，它们都具有由退火引物决定的固定的 5ˊ端以及终止于某一ddNTP的不定的 3ˊ端。

图 4.2 中介绍了两种双脱氧测序的工作方案。最早期的双脱氧法，本章称之为 Sanger 法，是利用大肠杆菌 DNA 聚合酶 I 大片段(或称 Klenow 片段，Klenow 酶)发展起来的。“ 标记／终止法 ”则利用了一种修饰的 T7DNA 聚合酶，在两个独立的反应中分别进行引物的标记和双脱氧核苷酸的掺入终止。引物与模板退火后，标记反应发生在 4 种低浓度 dNTP(其中 1 种是放射性标记)中，DNA 的合成持续到一种或多种 dNTP 被耗竭为止，这样可保证掺入全部的标记的脱氧核糖核苷酸。链终止反应在 4 个独立的反应中进行，每个反应除了含有 4 种 dNTP 外，还各含 4 种 ddNTP 中的一种,而高浓度的 dNTP 保证 DNA 逐次合成至生长链因 ddNTP 的掺入而终止。

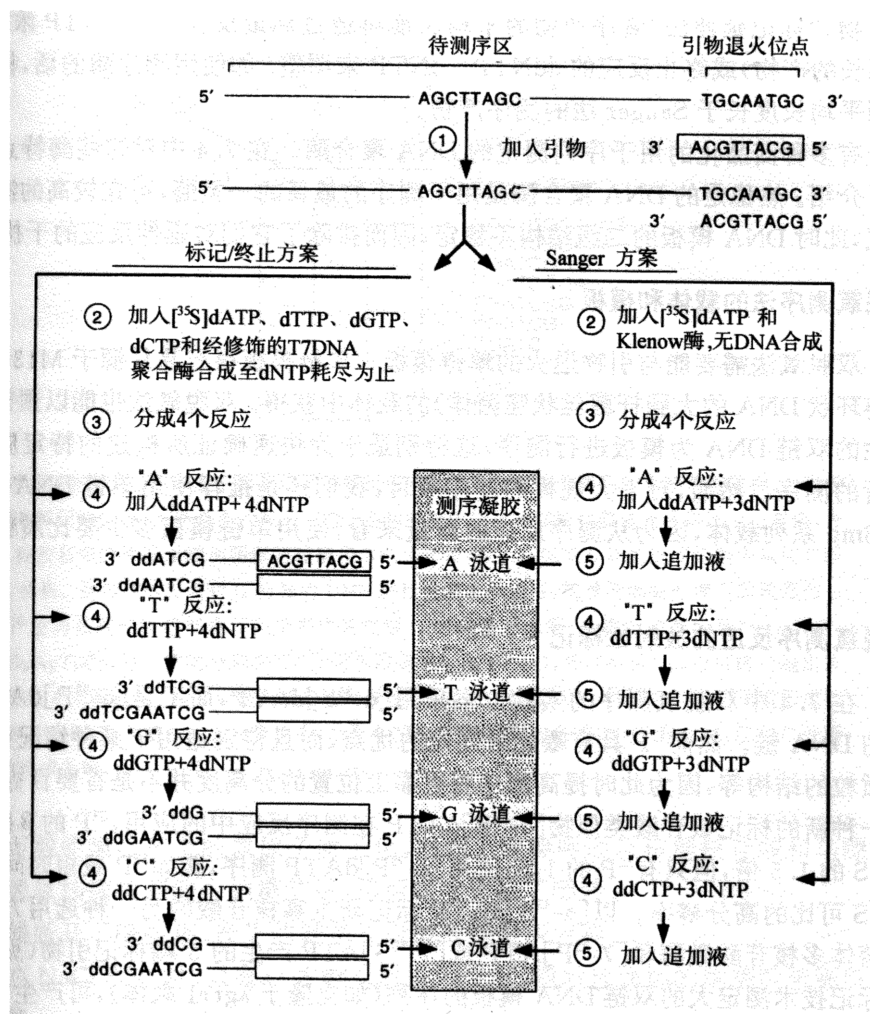


图 4.2 双脱氧测序法。在图示的每种方法中，单链 DNA 片段与引物退火后进行聚合反应(步骤 1),在 Sanger 法中((右图),加入 Klenow 酶和放射标记的 dATP(步骤 2),然后,分成 4 份进行反应(步骤 3),分别加入其余的 3 种 dNTP 和加入 ddATP、ddTTP、ddGTP 和 ddCTP 其中的一种(步骤 4)。DNA 合的成进行至摄入 ddNTP 后被终止。追加 dNTP(步骤 5)使未被终止的链再延伸以产生更高分子量的 DNA。“ 标记／终止法 ”(左图)说明略。在每种方法中，反应终止后，样品加样于测序胶的相邻泳道上，进行电泳分离（奥斯伯等，1998)。

Sanger 法测序产物的平均链长取决于 ddNTP：dNTP 的比例，比例高时，得到较短的产物；“标记／终止法”测序产物的平均长度可通过标记反应中 dNTP 浓度(高浓度能得到长的产物)或终止反应的 ddNTP:dNTP 来调整。

有多种商品化的用于序列测定的 DNA 聚合酶。热稳定的 DNA 聚合酶是用于测序的最新的一类酶，可在高的温度进行测序反应。此时 DNA 模板的二级结构不稳定，因而排除了它们对延伸反应的干扰。

3．化学测序法(Maxam-Gilbert 法)

在A．Maxam和W．Gilbert(1977)发展的DNA化学测序法中，与碱基发生专一性反应的化学试剂在一种或两种特定核苷酸位置上随机断裂已纯化的 3ˊ端或 5ˊ端标记DNA链，产生 4 套寡聚脱氧核糖核苷酸。在随后的测序胶放射自显影中，仅末端标记的片段显迹，故可得到如图 4.3 所示的 4 种DNA阶梯。

肼、硫酸二甲酯(DMS)或甲酸可以专一性地修饰 DNA 分子中的碱基，这构成了化学测序法的基础，加入吡啶可催化 DNA 链在这些被修饰核苷酸处断裂。化学法的特异性基于第 1 步反应中肼、硫酸二甲酯，或甲酸仅与 DNA 链上小部分特定碱基的作用，而第 2 步的哌啶断裂必须定量反应。第 1 步反应的化学机制如下：

G 反应： DMS 使鸟嘌呤的 7 位氮原子甲基化，其后断开第 8 位碳原子和第 9 位氮原子间的化学键，哌啶置换了被修饰鸟嘌呤与核糖的结合。

G+A 反应： 甲酸使嘌呤环上的氮原子质子化，削弱了腺嘌呤脱氧核糖核苷酸和鸟嘌呤脱氧核糖核苷酸中的糖苷键，然后哌啶置换了嘌呤。

T+C 反应： 肼断开了嘧啶环，产生的碱基片段能被哌啶所置换。

C 反应： 在 NaCl 存在时，只有 C 才能与肼发生反应，随后被修饰的胞嘧啶被哌啶置换。

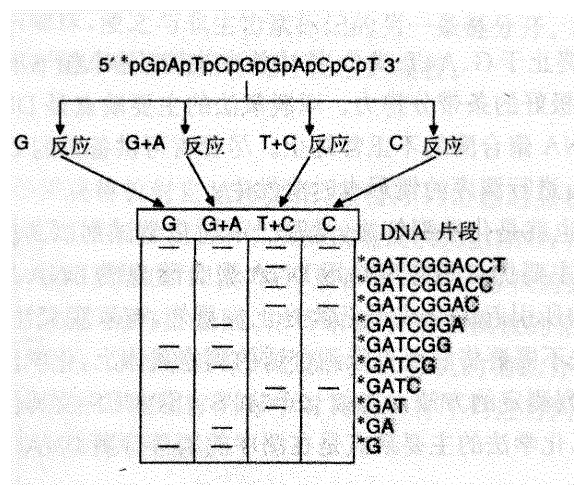


图 4.3 化学测序的策略。图中表示四个化学裂解反应产物经凝胶电泳分离后的寡核苷酸阶梯。“ * ”表示DNA片段上^{32}P标记的位置。本例是在片段的 5ˊ端。凝胶右侧的片段 3ˊ端加阴影的碱基表示经化学修饰后，在哌啶介导的链间切割中从核苷酸链上被取代的碱基（奥斯伯等，1998)。

4．荧光自动测序仪

自动化测序仪使凝胶电泳、DNA 条带检测和分析过程全部自动化。目前，所

有的商品化 DAN 自动化测序仪的设计都是以酶法(即 Sanger 法)测序反应产生荧光标记或放射性标记的测序产物为基础，它们都具有数据收集的能力，并含有进一步分析处理的程序。荧光标记物通过引物或 ddNTP 掺入到测序产物中。4 种碱基产生 4 种颜色的荧光反应，所以以单泳道或毛细管电泳就可以分辨出相应的寡核苷酸产物。

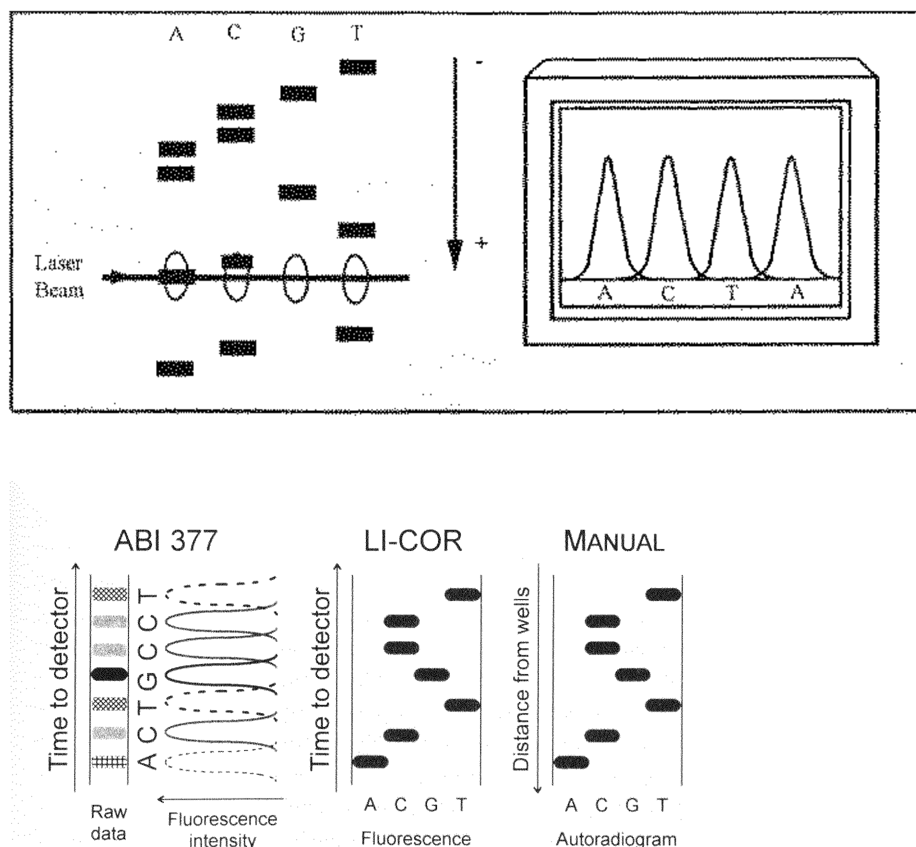


图 4.4 全自动测序仪基本操作原理

下面结合两种型号的 DNA 自动测序仪介绍自动测序原理。

ALF全自动激光荧光DNA测序系统（automated laser fluorescent DNA sequencer）是由德国海德堡（Heidelberg）欧洲分子生物学试验室（EMBL）W. Ansorge和B. Sproat提出和设计的。与同位素测序系统相比，ALF不但在仪器硬件设计上，而且在驱控仪器的软件功能上都作了很大改进。操作中能直接分析原始数据，也可以及时处理收集过程中获取的数据。最近推出的ALF express™全自动激光荧光核酸测序仪，则是利用电泳原理把荧光标记的DNA片段通过测序胶电泳分离。该仪器本身设计独特，提供快速可靠的核酸测序、片段分析、HLA序列定型及突变检测等。在人类基因组大规模序列测定中，该设备起到了重要的初筛作用。ALF express™系统采用非放射性的单一Gy5 荧光素标记引物或dNTPs进行核酸测序和片段分析，沿用Sanger双脱氧核酸末端终止测序法，使用Cy5 荧光标记的引物与模板进行退火。测试时，把A、C、G、T四种反应物分别加到凝胶板上的样品槽内，上样程序与手工测序相同。另外，在仪器电泳单元的下方是由激光枪(laser source)和探测器排列组成的探测系统：每个样品道后面都有一个探测

器，激光能透过凝胶的每一条泳道，当DNA条带迁移到探测区域并遇上激光时，DNA上的荧光标记立刻被激活，放出光信号；此荧光信号由泳道前的光探测器接收，并将信息输送给电脑进行分析和保存(图 4.4)。电泳结束后，电脑将收集到的信号(原始数据)进行处理，从而获得最终序列。

早在 1987 年 Perkin Elmer(PE)Applied Biosystems 公司就推出 DNA 自动测序仪，其专利是分别采用 4 种荧光染料进行标记且在同一个泳道测序，具有极大的优越性。377 型全自动 DNA 测序仪是 PE 公司近年推出的新型测序仪，它采用专利的四种荧光染料标记，并采用激光检测方法，具有测序精确度高、每个样品判读序列长(700bp)、一次电泳可测定样品数量多(64 个)、不需要同位素测序，方法灵活多样等特点，在人类基因组测序和 cDNA 文库测序研究中应用极其广泛。此外，该仪器在各种应用软件的辅助下还可以进行 DNA 片段大小分析和定量分析，应用于基因突变分析 SSCP、DNA 指纹图谱分析、基因连锁图谱表达水平的研究，有着极其广泛的应用前景。其原理是采用四种荧光染料标记终止物 ddNTP 或引物，经 Sanger 测序反应后，产物 3′端(标记终止物 ddNTP 法)或 5′端(标记引物法)带有不同荧光标记，一个样品的 4 个测序可以在一个泳道内电泳，从而降低了测序泳道间迁移率差异对精确性的影响。由于增加了一个电泳样品的数目，可一次测定 64 个或更多样品。经电泳后各个荧光谱带分开，同时激光检测器同步扫描，激发出的荧光经光栅分光后打到 CCD 摄像机上同步成像。也就是代表不同碱基信息的不同颜色荧光经光栅分光，经 CCD 成像，因而一次扫描可检测出多种荧光，传入电脑。其测序速度高达 200bp/h，比 373 型 DNA 测序仪速度大大提高。最后经过软件分析后输出结果。

自动化测序仪的发明促进了人类基因组的大规模测序行动。自动化测序效率高，而且测序的质量也比手工操作好。由于 DNA 多聚酶和荧光底物的不断更新，在很长一段时间内，荧光自动化测序将会处于主导地位。

二．DNA片段测序策略[2]

1．鸟枪测序法(shotgun sequencing)

大分子 DNA 被随机地“敲碎”成许多小片段，收集这些随机小片段并将它们全部连接到合适的测序载体；小片段测序完成后，根据重叠区计算机将小片段整合出大分子 DNA 序列。这就是所谓的鸟枪测序法(见图 4.6)。鸟枪测序法可以迅速获得 90%左右的片段序列结果，但随后测序效率明显下降，这是因为随后测定的随机片段越来越多地是重复已测序完成的片段。因此，一般通过合成特定的寡核苷酸引物来测定剩余少量未知片段。

有三种方法可用来将DNA大片段切割成小片段：限制性内切酶、超声波处理和DNA酶I降解(加Mn^{2+})。在这三种方法处理前， DNA的纯化非常重要，要去除载体DNA或仅由载体DNA产生的片段。

[2]本部分内容译自Alphey L. DNA Sequencing—From Experimental Methods to Bioinformatics，BIOS Scientific Publishes Limited ，1997

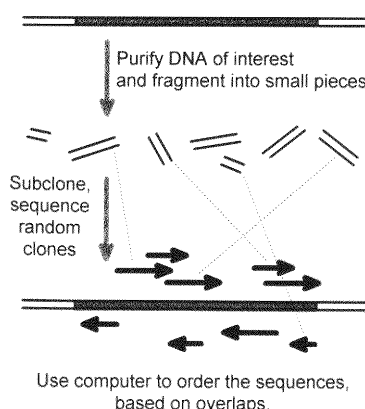


图 4.6 鸟枪法测序过程

鸟枪测序法的优点是成本低、快速、易于自动化操作，它的缺点是在测序后期，大量重复测序使测序效率变低。

1995 年第一个细胞有机体——流感嗜血(*Haemophilus influenzae*)全基因组序列被完成，这是完全用鸟枪法策略直接完成的，说明鸟枪法用于微生物基因组测序是有效的。研究者直接将全基因组 DNA 打成 1.6～2.0kb 大小的片段分别克隆，共使用了 19687 个模板，进行了 28443 个测序反应，组建了 140 个片段重叠群，测序用时 3～4 个月，耗费 100 万美金左右。

2．引物步查法(Primer walking)

引物步查法是一种渐进式测序策略，也是最简单的一种测序策略。该方法适合于双脱氧测序，并绕开了亚克隆小片段DNA的要求。最初的序列数据是通过利用载体上的引物获得的，一旦新的序列被确认，与新获得序列的 3'端杂交的寡核苷酸就能合成，并能以之为引物进行下一轮的双脱氧测序反应。这样，从两头向中间，序列被一步步测序(见图 4.7)

引物步查法相对较慢，因为序列仅从两头测得。每一步均需要一个测序反应(凝胶电泳)、数据分析、新引物设计和合成。这些过程将至少需要几天时间，如果引物供应不畅，可能时间还要更长。该方法适合于短 cDNA 片段，不适合于长 cDNA 片段，同时不宜自动化处理，因为每一反应需要一个不同的引物，这些引物将依据上一次反应结果而定。引物步查法成本相对较高，每一步都需要合成一个新引物，这制约了该技术的广泛应用。但是，最近寡核苷酸合成的成本已显著下降，所以成本问题有望解决。该技术的优点在于它的简单，不需要亚克隆或其它一些操作，实际操作时间不多，在其测序过程中，分析者有大量时间可以干其它一些事情。

引物步查法将合成一套覆盖整条序列的测序引物，如果序列需要重复测序，如测定序列突变等位位点，这套引物则成为很有用的资源。

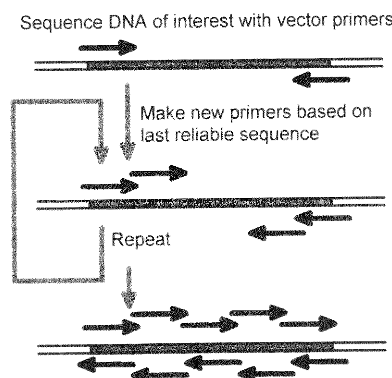


图 4.7 引物步查法测序过程

3．限制性酶切—亚克隆法(Restriction endonuclease digestion and subcloning)

原理上讲，序列的信息可以从其已知的限制性内切酶位点中获得。用限制性内切酶酶切并亚克隆一个适当大小的片段，使酶切位点附近的未知片段与载体已知序列相邻，这样就可以用载体的引物去测定未知序列；可以很方便地利用 2 个或更多位点切除一个未知克隆片段并用 DNA 聚合酶再将酶切下来的克隆产物再接合上去。由于所选用的内切酶不可能产生粘性末端，所以正常情况下，有必要用 Klenow 或 T4DNA 聚合酶把它们转变为平端。该方法示意图见图 4.8.

该方法的关键一步是需要一张准确的限制性内切酶谱，而且这些酶切位点间最好都相隔几百个碱基。对于一个熟练的研究者来说，制作一张酶切图并不难，但是酶切位点的分布则是一个随机问题，所以，不可能位点距离总是符合该方法的测序。利用该方法可以得到整条片段的大部分序列。由于该方法是基于酶切图，所以对于尚有哪些缺口(gap)，缺口有多大都很清楚，这有助于进一步的分析。

该方法难以自动化分析，因为它依赖于一套特定的亚克隆过程，而这些过程在每次的测序计划均是不同的。可能最常用的方法是用未知片段中的少量酶切位点，每个位点作为未知片段的一个新起点，然后用引物步查法在每个方向进行测序。这种混合方法较单用引物步查法可以显著减少整个片段的测序时间。

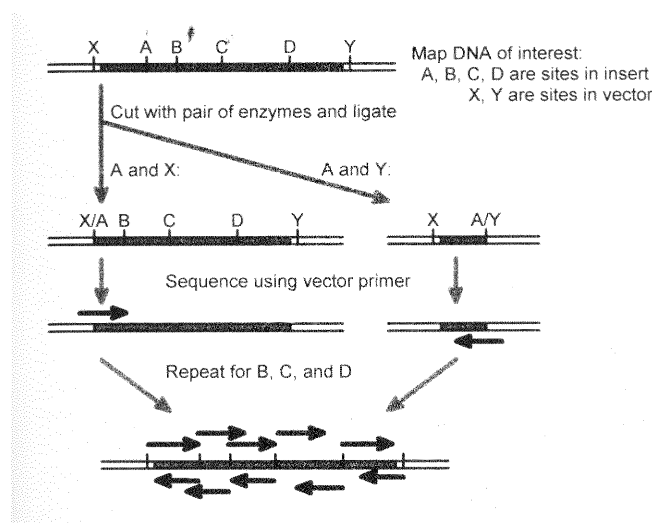


图 4.8 限制性酶切—亚克隆法测序过程

三、基因组测序策略

1. 逐步克隆（clone by clone）：从遗传图谱、物理图谱到基因组图谱[3]

基因组测序涉及 DNA 的大规模测序，它是一项如同“ 曼哈顿登月计划 ”一样的庞大工程，是人类在现有技术水平的重重障碍中科学技术的又一次进步。根据现有的技术水平，人类还无法对基因组这样的复杂 DNA 大分子直接进行测序，而只能采取分而治之的测序基本策略，即将基因组 DNA 分割成一定大小的片段，然后分别对这些片段进行测序。这样便产生了这样一个问题：如何将这些片段准确地拼接起来？目前的测序方法(上节)每次反应只能测定 500bp 左右长度的 DNA 片段，而一般一条染色体的长度对于 400-500bp 长度如同天文数字。所以，要进行诸如人类基因组测序，则必须在 2 个方面取得突破：一是将基因组 DNA 大分子分割并构建适合于测序的 DNA 片段库，而且库中的片段要覆盖整条序列；二是在整条线性序列上建立一定数量的“ 路标 ”，使切割下来的 DAN 片段能准确拼装回去。遗传图谱和物理图谱便是这样的“ 路标 ”图。人类遗传和物理图谱于 1998 年的建成使最终人类基因组测序成为可能。

基因组上的 DNA 相当稳定，因此可以构建含有这些 DNA 片段的新生物体。克隆技术是把基因组上的片段插入不同生物载体，并转染到一些生物体中使其生存和稳定复制，由此可以分析由小片段 DNA 组成的基因组拷贝(克隆群)。目前选用插入的载体包括酵母、细菌、粘粒、噬菌体等。

遗传图谱(genetic map)又称连锁图谱(linkage map)或遗传连锁图谱(genetic linkage map)，是指基因组内基因和专一的多态性DNA标记(marker)相对位置的图谱，其研究经历了从经典的基因连锁图谱到现代的DNA标记连锁图谱的过程。构建遗传图谱的基本原理是真核生物遗传过程中会发生碱数分裂，此过程中染色体要进行重组和交换，这种重组和交换的概率会随着染色体上任意两点间相对距离的远近而发生相应的变化。根据概率大小，人们就可以推断出同一条染色体上两点间的相对距离和位置关系。正因为如此，我们得到的这张图谱也就只能显示标记之间的相对距离。我们称这一距离(概率)为遗传距离(cM)，由此构建的图谱也称为遗传图谱。遗传图谱的‘ 路标 ”(遗传标记)已经历了几次从“ 粗 ”到“ 细 ”的大的演变，或者说，从第 1 代标记向第 2 代、第 3 代标记的过渡。经典的遗传标记(第 1 代标记)最初主要是利用蛋白质或免疫学等的标记，70 年代中后期建立起来的限制性片段长度多态性(RFLP)方法成为第 1 代的DNA标记，这类标记在整个基因组中确定的位点数目可达 10^5以上。第 2 代标记为可变数量串联重复序列(Variable number tandem repeat,VNTR)，包括微、小卫星(microsatellite/minisatellite)或短串联重复(short tandem repeat,STR或short sequent length polymorphysm,SSLP)标记等。第 3 代标记是一类称作SNP(single nuleotide polymorphysm)的遗传标记系统，即单核苷酸多态性标记。

遗传图谱上的各种DNA标记正如地图上标明的河流、山川，基因组中的这些标记种类繁多，随着人类基因组等计划的进行，人们不断发现一些新的标记，而且这些标记在地图上的密度也越来越高，迄今已经有好几个版本的图谱发表出来。在Internet网上的GDB(geneome database)网页上可以方便地查找到迄今已

[3]本部分内容取自陈竺、杨焕明等人的文章，见：贺林．解码生命—人类基因组计划和后基因组计划，北京：科学出版社，2000

发表的各种遗传标记(http://gdbwww.gdb.org)。

遗传图谱的构建是人类基因组研究必不可少的一步，它对搞清基因的功能、定位及分离克隆新基因、排列 DNA 片段、研究染色体上基因的排列顺序等起到不可估量的作用。遗传图谱在过去几年的人类基因组研究中发挥了巨大的作用，以致同样的策略也被应用于其它模式生物。

物理图谱是描述位于染色体上的基因和生物学界标独特并有确定位置及实际距离的染色体结构。任何图谱都是一系列路标及客观物(objects)按其固有的顺序和可能的距离构建出来的。客观物的顺序应不随构图方法的不同而不同，但它们之间的距离则可能不一致。在遗传图谱中按重组率来估计实际距离会有很大的偏差。物理图谱可以理解为用物理学方法而不是遗传学方法定位的由客观物组成的任何图谱，而通常物理图谱是指高分辨率(high-resolution)的物理图谱，即基因组长片段限制性酶切图谱和重叠克隆图谱等，但整合物理图谱还应包括只能粗略分辨路标位置但不能准确排位的染色体图谱(chromosome map)和遗传连锁图谱。

人类基因组测序的开展还得益于另一项突破：随着脉冲场电泳技术(pulsed-field gel electrophoresis,PFGE)、YAC 克隆、BAC 和 PAC 克隆的出现，可以把切割基因组后产生的大片段 DNA 准确地分离和纯化，并插入能转入 DNA 大片段的载体，转染酵母细胞形成 YAC 克隆库或转染大肠杆菌形成 BAC 克隆库。这些载体可载入 10Mb 长度(相当于人类全基因组碱基长度的 1/300)的 DNA 片段。全基因组的 YAC 克隆库及 BAC 克隆库保证了基因组分析的完整性和准确性。可以用杂交技术等来发现重叠克隆，以此进行克隆片段的排序。对于大片段 DNA 克隆进行再切割，并载入粘粒、细菌或噬菌体，即可构建相应于特定 YAC 或 BAC 克隆的亚克隆(subcloning)，供测序使用。这一系统过程的建立为大规模测序打下了坚实的基础。

构建物理图谱最终是要统一到基于 STS 的物理图谱。STS(sequence-tagged site,序列标签位点)的概念首先由 Olson 于 1989 年提出，目的是建立一套人类基因组统一的生物学界标。STS 本身是随机地从人类基因组上选择出来的长度在 200～300bp 左右的特异性短序列。STS 路标的建立一般是从噬菌体 M13 上构建特定染色体克隆开始，STS 概念的提出是物理构图的一次革命，由于特定 STS 在一套基因组结构中只出现一次，统一地把相应的克隆库中的克隆进行排序变得更准确和更科学。如果两个或两个以上的克隆包含有相同的 STS，则它们之间存在重叠。基于 STS 的物理图谱的重要性在于(1)它们可用来特异地定义 YAC、粘粒或噬菌体克隆；(2)STS 可鉴定出与特定克隆存在重叠的克隆；(3)在计算机数据库中的各种物理图谱可以用 STS 这种通用语言统一起来。基于 STS 的物理图谱不但可对染色体图谱、限制性酶切位点为路标的限制性酶切图、重叠探针杂交的 YAC 克隆片段重叠群(contig)图谱及其亚克隆重叠排序，以及新近发展的其它新方法构建的物理图谱进行整合，也可对遗传图谱、基因图谱等各类图谱进行整合，最终完成系统、统一的基因组终极图谱。最终完成的人类基因组核苷酸序列相当于 STS 密度最高的基因组物理图谱。

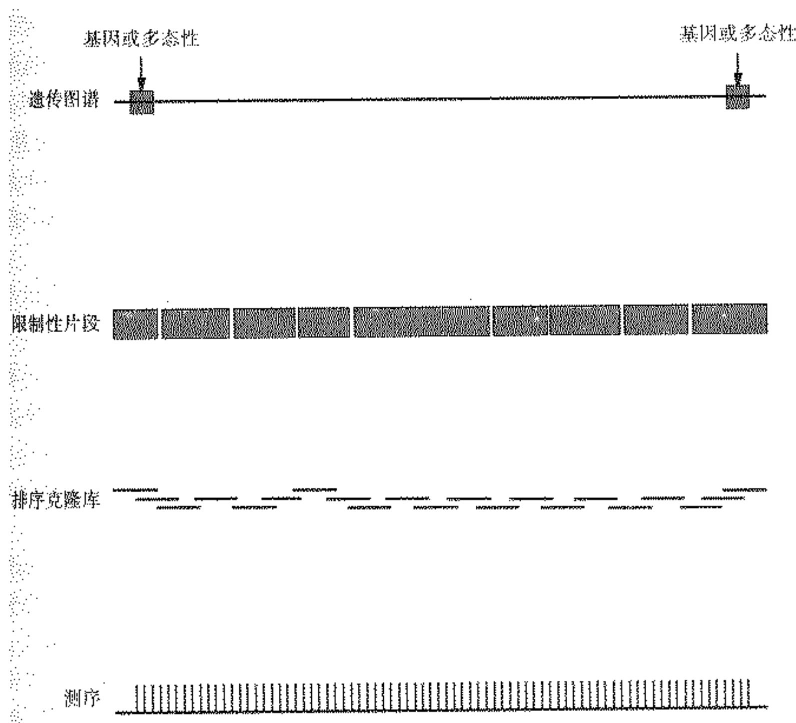


图 4.5 人类基因组的各种图谱。最粗糙的图谱是遗传图谱，它根据相邻标记（如基因和多态片段）间的重组率来测量相互间的距离；具有 1-2Mb 长度的限制性酶切片段可被分离和构建物理图谱；YAC 等长度在 40-400kb 的插入片段排列构建高分辨率物理图谱；碱基序列为最高分辨率物理图谱。

综上所述，广义上各种基于路标位点构建的物理图谱方法从低分辨率到高分辨率可主要分为以下几种：

(1)对路标进行粗略定位的染色体图谱即细胞遗传图谱(cytogenetic map)，通常使用原位杂交(ISH)或荧光原位杂交(FISH)技术确定含有路标 DNA 片段在染色体上的区带位置和分布。DNA 片段可被定在 2～10Mb 的范围内。

(2)cDNA 图谱是在细胞遗传图谱上显示 cDNA 或 ESTs(expressed sequence tags)，即表达 DNA(外因子)的区带位置。部分 cDNA 序列可作为路标。

(3)利用家系分离分析法(pedigree segregate analysis)可确定具有多态性的遗传标记位点在遗传连锁图谱上的位置，最新的人类基因组遗传连锁图谱已把标记间的平均距离缩小到 1cM 以下，即粗略地对应于物理图谱中的 1Mb 范围内。

(4)辐射杂种图谱是利用体细胞遗传技术(somatic cell genetic approach)构建高分辨率、长范围连续的人类基因组图谱。基本原理为，人为地用放射线打断染色体，制备出含有特定人类染色体或片段的杂交细胞系，并利用类似于传统的减数分裂构图原理确定路标间的距离和位置，最高的分辨率可达到 50kp。

(5)脉冲场电泳的长片段限制性位点(macrorestrietional site)图谱，即限制性酶切位点指纹(restriction enzyme fingerpriting)图谱是描述以稀有酶切位点为生物学界标的顺序和距离，以及形成基因组或染色体区域上的酶切图谱。由于些法是从大片段入手，常常又称为“ 从上到下 ”(Top-down)构图法；此外，区域性 DNA 大片段有利于较精细制图，如 YAC 克隆插入片段分析便于重叠图谱的分析，此方法可把 DNA 片段定位在 100kb 到 1Mb 范围内。

(6)由 DNA 片段重叠群(contig)形成的小组合，即相连组合图谱，或称重叠克隆群(overlapping sets of cloning)图谱描述存在于重叠的 DNA 片段克隆的顺序和距离。通常通过粘粒重叠克隆把 DNA 片段定位在小于 2Mb 的范围内，相对于长片段限制性酶切位点图谱，这种构图法也被称为“从下到上”(Bottom-up)法。

(7)序列标签位点(sequence-tagged site,STS)构成了 STS 基础上整合图。它是基因组上筛选特异序列，其最终密度至少达到平均每 100kb 左右一个，最终将把各种方法构建的图谱整合起来，完成准确完整的系统物理图谱。

(8)部分及全基因组测序是分辨率最高的物理图谱，而目前要构建的高分辨率(<100kb)物理图谱上路标序列本身也是基因组序列信息的一部分。

此外，一些构建物理图谱的方法还包括基因组序列抽样(genomic sequence sampling,GSS)和可见图谱(optical map)等。GSS 是结合片段限制性酶切和 STS 的一种作图法，分辨率可达到 1～5kb；可见图谱则是结合限制性酶切、电泳和 FISH 技术通过观察单个 DNA 大分子在限制性酶切作用下的图象来作图。

低分辨率物理图谱在人类基因组计划中本身是独立的部分，但从染色体区带－表达基因区域－遗传学距离－物理学实际距离－碱基序列这一过程来看，低分辨率染色体分带可看作粗略的物理图谱，碱基序列则是最精密的物理图谱。低分辨率图谱上的一些路标常常被用在高分辨率图谱的构建中，结合其它路标形成高密度路标分布的图谱，同时这些高密度路标可以重新在低分辩率图谱进行验证，形成高分辩率与低分辩率相结合的整合物理图谱。每种图谱都有各自的优缺点，所以即使对同一基因组研究，不同的实验室会采用不同的作图方法，但最终各种图谱的结果应能统一起来，相互补充和完善。

表 4.1 部分物种基因大小和遗传/物理距离关系

物种	拉丁名（英文名）	基因组大小（kb）	物理距离（kb/cm）
水稻	Oryza sativa (rice)	4.30×10^5	300
玉米	Zea mays (maize)	2.5×10^6	2140
小麦	Triticum aestivum (wheat)	1.6×10^7	
大麦	Hordeum vulgare (barley)	5.0×10^6	
燕麦	Avena sativa (oat)	1.1×10^7	
大豆	Glycine max (soybean)	1.2×10^6	
高粱	Sorghum bicolor (sorghum)	7.50×10^5	
马铃薯	Solanum tuberosum (potato)	8.4×10^5	
油菜	Brassica napus (rape)	1.1×10^6	
陆地棉	Gossypium hirsutum (upland cotton)	2.1×10^6	
黑麦	Secale cereale (rye)	9.1×10^6	
甜菜	Beta vulgaris subsp. Vulgaris (beet)	7.58×10^5	1100
西红柿	Lycopersicon esculentum (tomato)	9.5×10^5	510
拟南芥	Arabidopsis thaliana (thale cress)	1.20×10^5	139
洋葱	Allium cepa (onion)	1.5×10^7	
向日葵	Helianthus annuus (sunflower)	3.0×10^6	
菜豆	Phaseolus vulgaris (kidney bean)	6.3×10^5	

人	Homo sapiens	3.3×10^{6}	1000
小鼠	Mus musculus (mouse)	2.5×10^{6}	1800
大鼠	Rattus norvegicus(rat)	2.75×10^{6}	
线虫	Caenorhabditis elegans	9.7×10^{4}	250
果蝇	Drosophila melanogaster	1.37×10^{5}	500
大肠杆菌	Escherichia coli	4.6×10^{3}	
酵母	Saccharomyces cerevisiae	1.21×10^{4}	4.8
流感嗜血杆菌	Haemophilus influenzae	1.8×10^{3}	

表 4.2　基因组物理图谱数据库的部分相关网站。最新情况见附件。

数据库	网址
图谱相关资料	
STS 数据库(dbSTS)	http://www.ncbi.nlm.nih.gov/dbSTS/index.html
EST 数据库(dbEST)	http://www.ncbi.nlm.nih.gov/dbEST/index.html
CpG 数据库 (CpG island database)	http://biomaster.uio.no/CpGdb.html
细胞遗传图谱	
GDG	http://gdbwww.gdb.org
MGD	http://www.informatics.jax.org/mgol.html
辐射杂种图谱	
RHdb	http://www.ebi.ac.uk/RHdb
克隆重叠图谱	
YAC 克隆	
CEPH-GENETHON 整合图谱	http://www.cephb.fr/ceph-genethon-map.html
STS/YAC MAP	http://www_genome.wi.mit.edu/
BAC 和 PAC 克隆	http://www.tree.caltech.edu/
粘粒克隆	http://gea.lif.icnet.uk/
整合图谱	http://cedar.genetics.soton.ac.uk/public_html

表 4.2 中列举的物理图谱数据库的数据主要来自人类基因组，但同时也包含了其它的一些生物体。

构成物理图谱的 4 个基本要素之一可复制 DNA 片段(clonable fragment)(另 3 个要素是路标、单位、顺序)主要包括辐射杂种细胞(RH)、YAC、BAC、PAC 等。对于这些 DNA 大片段的测序一般需要将其再细分为能单独进行序列分析的小片段，目前有三种常用方法：鸟枪测序法、引物步查法和限制性酶切—亚克隆法。

2．全基因组鸟枪法（whole-genome shortgun）

在基因组水平上，全基因组鸟枪法和逐步克隆测定法是目前广泛应用的两个测序策略。小的单分子基因组，如细菌和小基因组(<10Mb)可直接用鸟枪法测序。虽然有人提出用鸟枪法直接测序人类基因组(Weber 和 Mayers，1997)，但由于人类基因组中存在高比例的重复序列(尤其是 LINE，2-7kb)、克隆文库不可避免的间隙和基因的多态性等原因，鸟枪法的片段组装几乎是不可能的。受读序长度的限制，一个反应无法跨过 LINE。鸟枪法在小组因组(1-5Mb)测序方面已取得了非常好的效果，例如流感嗜血杆菌(*H. influenzae*，1.9Mb)、枝原体(*M.genitalium*，0.58Mb)和甲烷球菌(*M.jannaschii*)基因组均用此法完成测序。逐步克隆测定法则通过建立克隆文库(YAC、BAC、PAC、Cosmid、Fosmicl、噬菌体、质粒)，然后用鸟枪法进行克隆片段的测序。所以，大规模测序的两个前沿基本都是采用鸟枪法(图 4.9)。

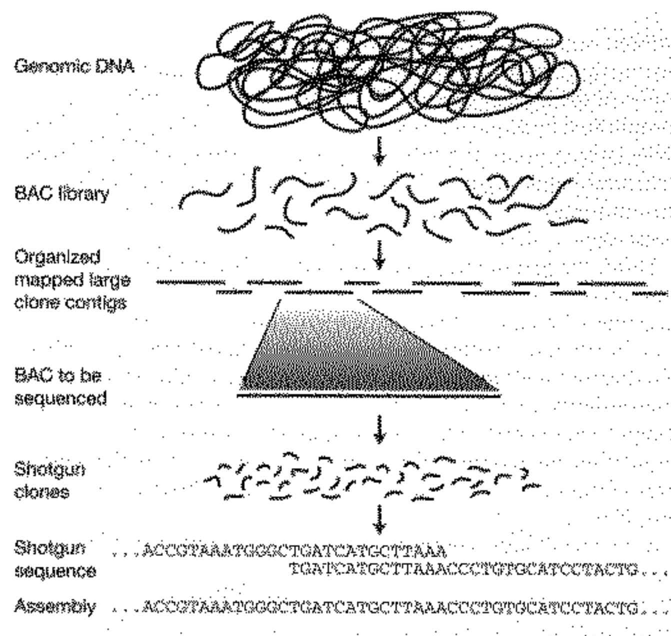


图 4.9 鸟枪法测序策略。基因组的逐步克隆测序包括图中的所有步骤：DNA 单链→构建 BAC 文库→鸟枪法克隆测序→组装；全基因组鸟枪法测序则省去中间的构建 BAC 文库步骤。

四、序列片段的拼接方法

无论是逐步克隆测序还是全基因组鸟枪法测序，都存在片段拼接组装的难题。目前 DNA 自动测序仪每个反应只能测序 500bp 左右，如何将这些片段拼接成完整的 DNA 序列呢？Lander 和 Waterman(1988)提出利用"指纹"(fingerprinting)随机克隆进行基因组作图的算法，它为大量鸟枪法随机测序的片段用计算机进行自动拼接提供了可能。这种技术不仅避免了传统的亚克隆策略的大量繁琐工作，还使测序具有一定的冗余性(即一定数量的重复)，保证了测序中每个碱基的准确性。

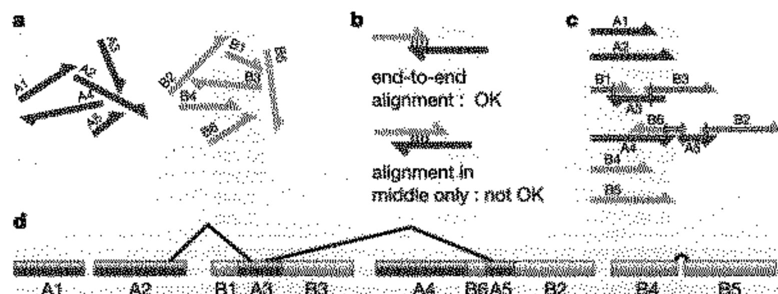


图 4.10 序列拼接示意图。a-d 步骤表示从单一克隆片段拼装成基因组草图的过程；A1-A5 和 B1-B6 分别表示由鸟枪法克隆 A 和 B 获得的序列重叠群。

目前 DNA 序列拼接应用的主要软件是由美国华盛顿大学 Phil Green 实验室

开发的 Phred-Phrap-Consed 系统。Green 也因研制该系统而在人类基因组研究历史上占有一席之地(见 *Science* 2001 年 2 月 16 日人类基因组专刊"A history of Human Genome Project"一文)。Phred(测序器)是一种碱基识别系统(base-caller),它根据自动测序仪信号按顺序识别碱基,估计测序错误率等。Phrap(组装器)是根据 Phred 的结果从头组装由鸟枪法产生的不同的短序列。Consed(校对器)与 Phrep 组成一个有机整体,利用 Phrap 组装的序列由 Consed 编辑、整合人工校对结果等。目前 36 个国家 900 多个实验室都在使用上述系统。非赢利研究机构或个人可申请免费利用该系统。

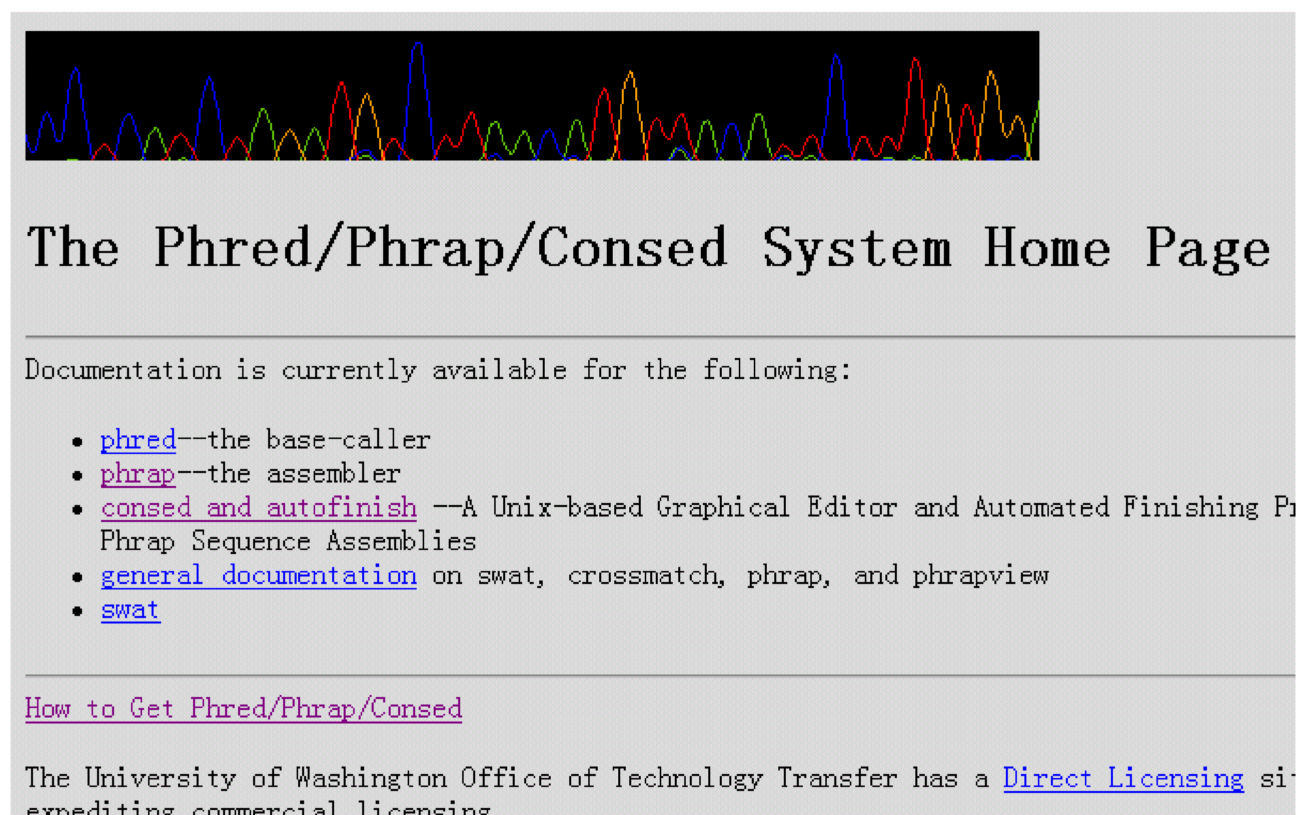


图 4.11 自动测序组装系统 Phred-Phrap-Consed 主页

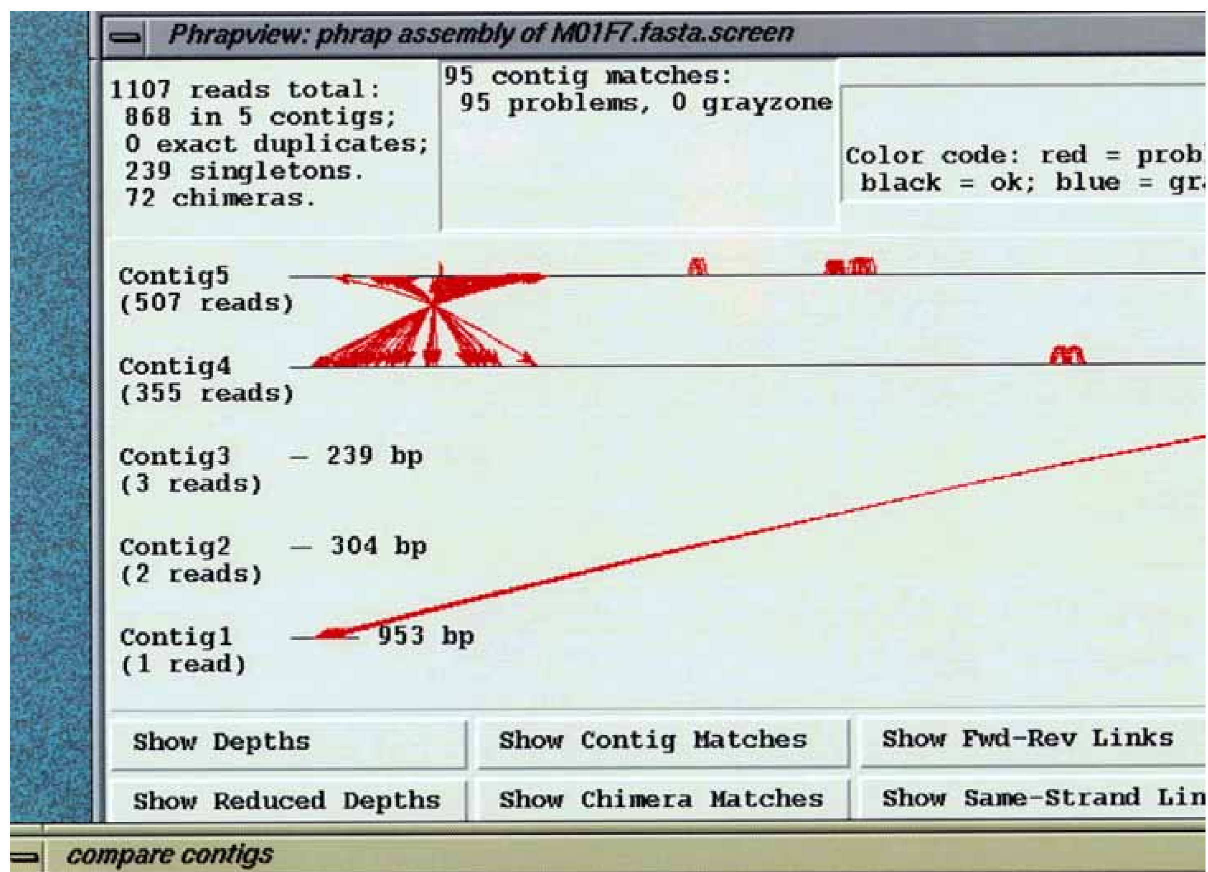


图 4.12 Phrap 程序中序列重叠群比对结果显示窗口

Phrap 拼接鸟枪法序列的的方法也是通过列线(aligment)查找匹配序列。其列线算法采用的是 Smith-Waterman 算法和 Needleman-Wunsch 算法（可选择），替换矩阵(缺省为 BLOSUM50)、空位设置罚值和空位扩展罚值(缺省值分别为-12 和-2)、E 值(缺省值 1.0)等都在列线比对中被应用。Phrap 的算法中使用了一个新参数——Z 值(Z-score)。当数据库序列长度变化很大时(实际情况往往如此)，理论分析和经验研究都表明列线值敏感性下降，即判别由随机性产生匹配的能力下降。Z 值的引入便是为了解决这一问题。Z 值定义如下：

$$Z = [s - f(n)] / \sqrt{g(n)}$$

其中 s 和 n 为原始列线值和数据库序列长度，f(n)和 g(n)分别是序列长度为 n 的序列列线值平均数和变异度。由此，Z 值的平均数为零，标准差为 1,与序列长度 n 无关。相对而言，Z 值与数据库大小无关，这一特性与原始列线值 s 相似，但与 E 值不同，所以，Z 值是比 s 值更合理的一个指标尺度。

五、EST 测序及其分析

（待补充）

第二节　　基因组注释：基因区域的预测

一．从序列中寻找基因

1．基因及基因区域预测

在完成序列的拼接后，我们得到的是很长的 DNA 序列，甚至可能是整个基因组的序列。这些序列中包含有许多未知的基因，将基因从这些序列中找出来是生物信息学的一个研究热点。

基因一词最早是由丹麦遗传学家约翰逊(Johannsen W.)于 1909 年提出，而在这之前，遗传学创始人孟德尔用“遗传因子”表达了对基因的朦胧认识。基因的概念随着遗传学、分子生物学等的发展而不断完善。从分子生物学角度看，基因是负载特定生物遗传信息的 DNA 分子片段，在一定条件下能够表达这种遗传信息，产生特定的生理功能。基因按其功能可分为结构基因和调控基因：结构基因可被转录形成 mRNA，并进而转译成多肽链；调控基因是指某些可调节控制结构基因表达的基因。在 DNA 链上，由蛋白质合成的起始密码开始，到终止密码子为止的一个连续编码序列称为一个开放阅读框(Open Reading Frame,ORF)。结构基因多含有插入序列，除了细菌和病毒的 DNA 中 ORF 是连续的，包括人类在内的真核生物的大部分结构基因为断裂基因，即其编码序列在 DNA 分子上是不连续的，或被插入序列隔开。断裂基因被转录成前体 mRNA，经过剪切过程，切除其中非编码序列(即内含子)，再将编码序列(即外显子)连接形成成熟 mRNA，并翻译成蛋白质。假基因是与功能性基因密切相关的 DNA 序列，但由于缺失、插入和无义突变失去阅读框而不能编码蛋白质产物。

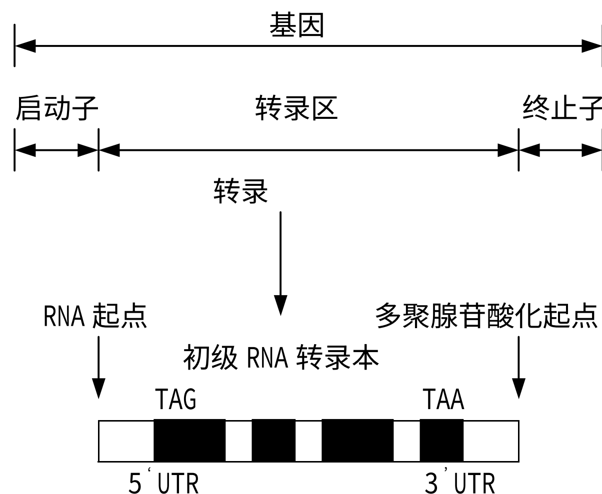


图 4.13 一种典型的真核蛋白质编码基因的结构示意图。其编码序列（外显子）是不连续的，被非编码区（内含子）隔断。

所谓基因区域预测，一般是指预测 DNA 序列中编码蛋白质的部分，即外显子部分。不过目前基因区域的预测已从单纯外显子预测发展到整个基因结构的预测。这些预测综合各种外显子预测的算法和人们对基因结构信号(如 TATA 盒等)的认识，预测出可能的完整基因。

某一算法的优劣可以通过一定的标准衡量：敏感性(sensitive)和特异性

(specifity)。假设待测序列中有M条序列是基因序列，而剩余的为非基因序列。我们用某一程序(算法)对待测序列进行预测，共预测出N条基因序列，而这N条序列中有N_1条确实为基因。则敏感性定义为N_1/M，它表示程序预测的功能；特异性定义为N_1/N，它表示程序预测结果的可靠程度。敏感性和特异性往往是一对矛盾。

基因区域的预测是一个活跃的研究领域，先后有一大批预测算法和相应程序被提出和应用，其中有的方法对编码序列的预测准确率高达 90%以上，而且在敏感性和特异性之间取得了很好的平衡。预测方法中，最早是通过序列核苷酸频率、密码子等特性进行预测(如最长 ORF 法等)，随着各类数据库的建立和完善，通过相似性列线比对也可以预测可能的基因。同时，一批新方法也被提了出来，如隐马尔可夫模型(Hidden Markov Model,HMM)、动态规划法(dynamic programming)、法则系统(ruled-based system)、语言学(linguistic)方法、线性判别分析(Linear Discriminant Analysis,LDA)、决策树(decision tree)、拼接列线(spliced alingment)、博利叶分析(Fourier analysis)等。

表 4.3 列出了 claverie(1997)对部分程序预测基因区域能力的比较结果，表中同时列出了相应算法和程序的网址。

目前基因区域预测的各种算法均基于已知基因序列。如相似性列线比较算法是完全依赖于已知的序列，而象 HMM 之类的算法都需要对已知的基因结构信号进行学习或训练，由于训练所用的序列毕竟是有限的，所以对那些与学习过的基因结构不太相似的基因，这些算法的预测效果就要大打折扣了。要解决以上问题，需要对基因结构进行更深入的研究，寻找隐藏在基因不同结构中的内在统计规律。

表 4.2 部分程序预测基因区域能力的比较结果（claverie，1997)

程序名称	所用算法	作者	预测对象	敏感性(%nucl)	物异性(%nucl)	敏感性(%exact exon)	特异(%exact exon)	丢失性的外显子(%)	错误的外显子(%)	网址
FGENEH	LDA	solovyev et al 1995	基因结构	83	93	73	78	15	11	http://dol.imgen.bcm.tmc.edu:9331/gene-inder/gf.html
GeneID	RB	Guigo et al 1992	基因结构	69	77	42	46	28	24	http://geneid@darwin.bu.edu www.imim.es/GeneIdentification/Geneid/geneid_input.html
GeneParser	DP	Snyder&Stormo 1993	基因结构	66	79	35	40	29	17	http://Beagle.colorado.edu/~eesnyder/GeneParser.html
Genie	HMM,DP	Henderson et al 1997	基因结构	87	88	69	70	10	15	http://www-hgc.lbl.gov/inf/genie.html
GenLang	LM	Dong&Searls 1994	基因结构	72	79	51	52	21	21	http://www.chil.upenn.edu/~sdong/genlang_home.html
GENSCAN	HMM,DP	Burge&Karlin 1997	基因结构	93	93	78	81	9	5	http://genomic.stanford.edu/GENSC-ANW.html
HEXON	LDA,DP	Solovyer et al 1994	基因外显子	88	80	71	65	10	27	http://dot.imgen.bcm.tmc.edu:9331/gene-finder/gf.html
MORGAN	DT	—	基因结构	83	79	58	51	14	-	http://www.cs.jhu.edu/labs/compbio/morgan.html
MZEF	—	Zhang 1997	基因外显子	87	95	78	86	14	7	http://Clio.cshl.org/genefinder
VEIL	HMM,DP	Krogh et al 1994	基因结构	83	72	53	49	19	-	http://www.cs.jhu.edu/labs/compbio/veil.html

注释：
①LDA：线性判别分析；RB：法则系统；DP：动态规划法；HMM：隐马尔可夫模型；DT：决策树；②敏感性(%nucl)：实际编码序列被成功预测为编码序列；特异性(%nucl)：预测为编码的序列实际确定为编码序列；敏感性(%exact exon)：实际的外显子被准确预测(包括拼接位点)；特异性(%exact exon)：预测为外显子的序列与实际外显子准确符合；丢失的外显子(%)：未能预测出的实际外显子；错误的外显子(%)：预测为外显子的序列实际不是任何外显子的片段。

2．发现基因的一般过程

从序列中发现基因可以理解为基因区域预测和基因功能预测 2 个层次。生物信息学在这 2 个层次上均形成具有自身学科特色的算法和手段，以下便简单描述通过生物信息学手段发现基因的一般过程。有关基因功能的预测将在以后的章节中进一步论述，同时本小节描述的发现过程只是生物信息学手段的一种可选策略。

以下主要根据 Gene Discovey(http://bioinformatics .weizmann.ac.il)：

第一步：获取 DNA 目标序列

① 如果你已有目标序列，可直接进入第 2 步；

② 可通过 PubMed 查找你感兴趣的资料；通过 GenBank 或 EMBL 等数据库查找目标序列。

第二步：查找 ORF 并将目标序列翻译成蛋白质序列

利用相应工具，如 ORF Finder、Gene feature(Baylor College of Medicine)、GenLang(University of Pennsylvania)等，查找 ORF 并将 DNA 序列翻译成蛋白质序列。

第三步：在数据库中进行序列搜索

可以利用 BLAST 进行 ORF 核苷酸序列和 ORF 翻译的蛋白质序列搜索。

第四步：进行目标序列与搜索得到的相似序列的整体列线(global alignment)

虽然第三步已进行局部列线(local alignment)分析，但整体列线有助于进一步加深目标序列的认识。

第五步：查找基因家族

进行多序列列线(multiple sequence alignment)和获得列线区段的可视信息。可分别在 AMAS(Oxford University) 和 BOXSHADE(ISREC,Switzerland)等服务器上进行。

第六步：查找目标序列中的特定模序

① 分别在 Procite、BLOCK、Motif 数据库进行 profile、模块(block)、模序(motif)检索；

② 对蛋白质序列进行统计分析和有关预测

第七步：预测目标序列结构

可以利用 PredictProtein(EMBL)、NNPREDICT(University of California)等预测目标序列的蛋白质二级结构。

第八步：获取相关蛋白质的功能信息

为了了解目标序列的功能，收集与目标序列和结构相似蛋白质的功能信息非常必要。可利用 PubMed 进行搜索。

第九步：把目标序列输入“提醒”服务器

如果有与目标序列相似的新序列数据输入数据库，提醒(alert)服务会向你发出通知。可选用 Sequence Alerting(EMBL)、Swiss-Shop(Switzerland)等服务器。

3．解读序列(making sense of the sequence)

在 2001 年二月份的第二星期里(12 日-18 日)，*Science* 和 *Nature* 同时刊发了具有划时代意义的人类基因组研究专刊。在 *Science* 的专刊中，有一篇题为“解读序列”(making sense of the sequence)(Galas D.J.)的综述文章。文章对序列，特别是人类基因组序列如何解读进行了深入分析，比较全面地展示了人类目前对序列的理解能力和技术现状。以下内容摘译自该篇文章。

利用基因组序列解决生物学问题已经具备了其自身(学科)特色，它被冠以“功能基因组学”。自从 1996 年酶母(*Sacharomyces cerevisiae*)基因组序列被公布，我们已熟悉用全基因组序列来研究基因表达模式等等生物学问题。虽然我们还不知道约 1/3 酶母基因的功能，但是我们知道所有与细胞功能有关的可能的蛋白质和 RNA 均由我们已知的序列编码。

根据目前对基因的分析结果，哺乳动物一个基因的转录产物平均有 2～3 种或者更多。从现有序列数据估计，人类的基因数约为 3 万，这意味着人类基因组编码了约有 9 万或更多种蛋白质。但是，以上由现有序列数据推测的结论有很多不确定因素。重叠序列群(contig)是由单个测序反应测得的序列(通常 400～800 碱基长度)拼装而成的一条连续片段，重叠序列群的数量和长度分布是基因分析的两个重要参数。正如美国 NCBI2000 年 12 月 12 日的报告所说，目前公共数据库中最大的重叠序列群为 28.5Mb，其中 43 个超过 1Mb，566 个在 250Kb～1Mb 之间，而 1628 个在 100～250Kb。这意味着长度大于 100Kb 的重叠序列群总长度约 600Mb——不足人类基因组全部序列的 20%；而基因组的一半序列是由 22Kb 或更小的重叠序列群所涵盖。由于基因的长度(一般估计为 30000 碱基对)大于或等于重叠序列群，这说明一定比例的人类基因不可能只在一个重叠群中；在一个重叠群中发现一个最长的基因，如肌联蛋白(Titin)基因(约 250Kb,内含 200 多个外显子)，比发现一个短的简单基因，如嗅感受蛋白基因(平均小于 2Kb)的概率小得多。但要将序列缺口和重叠群扩大还要籍以时日。因此，在不久的将来，基因的合成将通过组配重叠群“镶嵌物”(mosaic)，或称为“支架”(scaffold)来完成，这意味着重叠群间的拼接又将增加序列数据的不确定性。

要想将所有的基因都落入拼装而成的无缺口的支架片段中似乎还不可能，但是组装成的基因的大致轮廓将变得很清楚。这就象一个被重新复原的古希腊花瓶，虽然花瓶的残缺部分被用陶土填补，而整个花瓶的轮廓已很清晰。文特尔(Venter)等人进行基因拼装和分析的方法中，一人重要的参数是支架的大小和分布。据报道，支架的平均长度超过 1Mb，而 10Mb 以上的支架占整个基因组的 25%，支架间的缺口平均只有 2Kb。这些为基因分析者提供了高档次的序列数据。

从一给定序列片段中，通过相似性比较发现基因的效果决定于简单的统计量和重叠群在基因组中的覆盖率。当该覆盖率达到 90%以上，那就意味着几乎所有的基因(或至少是基因片段)均可在序列数据中找到。因此，利用本周公布的数据(指 *Science* 和 *Nature* 的人类基因组专刊)，通过相似性搜索来发现任何一个基因几乎都是可能的。

但是必须注意的是，这样确定的基因可能还具有随意性。这是因为某一生物，例如果蝇(*Drosophila*)的一条具有高度相似的受体基因序列可能来自几个不同的同源基因，而这些基因可能具有相同或完全不同的功能，甚至可能是一些没有功能的假基因(pseudoge)。也就是说，共同的功能域(domain)或模序(motif)可能在几个基因同时存在。使用 BLAST 搜索工具可能还是目前发现相似序列的最佳途径。NCBI 网站简明的介绍内容有助理解不断增多的 BLAST 系列工具的特性，有些小册子介绍了 BLAST 近似算法的统计特色和局限。BLAST 算法并不适合于所有目的的近似估计，但使用者应有这样的认识，即任何一种算法都有可能错过一些特殊相似性。例如，由于对一些相隔相似性(interrupted similarity)的忽略，使间隔越大，获得相似性统计显著的可能性越小。新的一些方法试图利用编码区的结构因素来进行相似性比对，这突破了相似显著性方法的局限。

虽然在基因组序列基因的自动化识别方面已取得巨大进步，但根据序列构建

准确的基因模型(model of genes)还需要大量的人力，即“手工操作”(“hand-on” effort)。基因的最佳模型是其全长 mRNA 序列。RNA 序列(以 cDNA 形式)可以将基因组序列基因的外显子结构串联起来，而不必考虑这些片段身处何方——片段的连续性、顺序和方向并不影响串联过程。但是，假基因和高度相同的重复序列可能使这一策略失灵，这引起了对收集更多全长 cDNA 序列数据的争论。

大致有 2 条途径可以发现基因：(1)基于同源性的方法，包括已知 mRNA 序列的应用；(2)基因家族和特殊序列间的比较。最初的方法包括利用各种计算机手段分析外显子和其它序列信号，如酶切位点等。

在每一个基因模型中，与调控相关的序列位置和结构往往是最难完成的注释(annotation)之一。在一些情况下，可以通过诸如模序(motif)(检索)来寻找和鉴定这些重要序列区段，但是我们目前对调控区段的鉴定和预测能力还很有限和不可靠。特定基因组间的比较是获得这些区段的一条途径，它建立在可以通过比较找出保守区的假设基础上。新的一些实验方法，例如列阵技术可以定位基因组水平的转录位点，同样可以有效地检测出基因组顺式调节(cis-regulatory)信号。

目前已有很多工具可以用于自动注释工作，对于这些工具的特点本文不做进一步论述。将统计学和启发式机器学习方法(heuristic methods) 相结合来分析基因和基因特征是目前流行的趋势(例如隐马尔可夫模型、神经网络和贝叶斯网络)。它们在发现基因方面最有效的方法并不是在准确建模方面，而是常与同源性方法配合使用。影响这些算法有效性的因素包括测序误差和统计偏差，例如碱基组成。数据的噪音(noise)会极大降低这些方法的效果，所以以上基于误差率较高的序列草图的预测结果将明显劣于基于完成序列的预测。

GENSCAN(http://genes.mit.edu/GENSCAN.html)是被广泛用于基因查寻和预测的软件之一，但是一些新软件，如 Genie 也不逊色。Genie (http://www_hgc.lbl.gov/inf/genie.html)是一种隐马尔可夫模型(HMM)系统，它可以整合不同来源的信息，如信号传感器(酶切位点、起始密码等)、内含子和外显子、mRNA EST 的列线和肽序列等。其它软件工具，如 GENEBUILDER、GLIMMERM、FGENES、GRAIL 等，最近也都被评价过。有一个简单的办法可以比较这些软件的优劣：利用果蝇基因组数据为例，GASPI 项目(Genome Annotation Assessment Project)(www.fruitfly.org/GASPI)对真核生物基因组注释的进展和存在的问题进行很好的比较分析。另外利用拟南芥(Arabidopsis)基因组也进行了相同的比较分析。

Nature 和 *Sciece* 上的两篇人类基因组分析论文分别使用了各自的基因分析系统。由公共资金资助的人类基因组计划(IHGSC)(论文发表在 *Nature* 上)使用的是一个称为“Ensembl”的系统，它使用 GENSCAN 进行初步预测，GENSCAN 利用 mRNA、EST 和蛋白质模序信息进行比对；然后使用 GeneWise(www.Sanger.ac.uk/software/Wise2/)进行蛋白质匹配分析，GeneWise 曾被用于果蝇基因组分析。以文达尔(Venter)为代表的私人公司(论文发表在 *Science* 上)使用的是一种称为“otto”的专家注释系统(rule-based expert system for annotation)，该系统力图将人的一些智能纳入程序中。

二、最长 ORF 法等：基于编码区特性

基因区域或蛋白质编码区的识别，特别是对高等真核生物基因组 DNA 序列中编码区的识别仍未能实现完全自动化。将每条链按 6 个读框全部翻译出来，然

后找出所有可能的不间断开放阅读框(ORF)往往有助于基因的发现。预测基因组的全部编码区或称为开放阅读框的方法概括来说也可以分为三类：一类是基于编码区所具有的独特信号，如始起密码子、终止密码子等；二是基于编码区的碱基组成不同于非编码区，这是由于蛋白质中 20 种氨基酸出现的概率、每种氨基酸的密码子兼并度和同一种氨基酸的兼并密码子使用频率不同等原因造成的；三是通过同源性比较搜寻蛋白质库或 dbEST 库寻找编码区。前二类方法主要是利用编码区的特性来寻找，本小节对这二类方法做简单描述。

最长 ORF 法：

在细菌基因组中,蛋白质编码基因从起始密码 ATG 到终止密码平均有 100bp，而 300bp 长度以上的 ORF 平均每 36Kb 才出现一次，所以只要找出序列中最长的 ORF(>300bp)就能相当准确地预测出基因。

在真核生物中，全长 cDNA 的编码区一般也可以用最长 ORF 法，如水稻的 3 万多条的全长 cDNA 的编码区预测（见 KOME DATABASE）。但是，要十分小心的是，这一预测有时也会出错。例如：以下全长 cDNA 的编码蛋白序列应为 4-029B，而非最长的 4-029A。

```
>4_029
ATCGGCCATTACGGCCGGGGACACAACAAACCAACAAACATCATAATTAACCTCTTCCTCCCAAGTAGT
CATCTGCCAACATGAAAGCCCTCGCACTCTTCTTCGTACTTTCCCTCTATCTCCTCGCCAACCCAGCTC
ATTCCAAGTTCAATCCCATCCGCCTCCGCCCCGCCCACGAAACGGCGTCGTCCGAAACTCCGGTGCTCG
ACATCAACGGCGACGAAGTCCGGGCCGGCGAAAATTACTACATTGTCTCCGCCATATGGGGCGCCGGCG
GAGGAGGCCTGAGACTCGTCCGATTGGATTCCTCCTCGAACGAATGCGCCAGCGACGTGATCGTATCCC
GGAGCGACTTCGACAACGGCGACCCGATTACCATCACGCCGGCGGACCCGGAATCCACCGTCGTCATGC
CGTCGACGTTCCAGACCTTCAGATTCAACATTGCGACCAACAAACTCTGCGTAAACAACGTAAACTGGG
GGATCAAGCACGACAGTGAATCCGGGCAATATTTCGTGAAAGCCGGCGAGTTCGTCTCCGACAATAGCA
ACCAGTTCAAGATTGAGGTGGTCAACGACAACCTTAACGCTTACAAAATCAGTTATTGTCAGTTCGGCA
CCGAGAAATGCTTCAACGTTGGCAGATACTACGACCCGTTGACCAGGGCTACGCGTTTGGCTCTCAGTA
ATACTCCCTTCGTGTTTGTGATCAAACCTACTGATATGTAATGAGCACCGGTGTTGAGGTTGCATGCAT
GTTATGGAGCTATGCTAAATAAGTAACGTTGCAACTTTGACAACGTTGTACGTGTAATAATAAGAATAA
ACATGCAATAAATCCGAGCTTGTTGTGTTGTGTAAATTTAACTATCTTAAATGAATAAGCATAATATTA
TCTATGCGAAAAAGAAAAAATAATAAAAAAAATTCATGTTCCGCCGCCTCGGCCCAGTCAACTCTGAAT
CCAAGCAAGCTTATGCATGCGGCCCAAATTCAAGCTCAATTGGCCAATTCGCCTATAGGGAGTCGTATT
ACATTCATGGCCGTCGTTTTACACGTCGGGACTGGGAAAACCCTGGGGTTACCCAACTTATCCCCTTGG
GCCCATTCCTCC

>4_029A ORF:69..755 Frame -2  Most length 687
MQPQHRCSLHISRFDHKHEGSITESQTRSPGQRVVVSANVEAFLGAELTITDFVSVKVVVDHLNLELVA
IVGDELAGFHEILPGFTVVLDPPVYVVYAEFVGRNVESEGLERRRHDDGGFRVRRRDGNRVAVVEVAPG
YDHVAGAFVRGGIQSDESQASSAGAPYGGDNVVIFAGPDFVAVDVEHRSFGRRRFVGGAEADGIELGMS
WVGEEIEGKYEEECEGFHVGR
>4_029B ORF:81..731 Frame +3 second length 651
MKALALFFVLSLYLLANPAHSKFNPIRLRPAHETASSETPVLDINGDEVRAGENYYIVSAIWGAGGGGL
RLVRLDSSSNECASDVIVSRSDFDNGDPITITPADPESTVVMPSTFQTFRFNIATNKLCVNNVNWGIKH
DSESGQYFVKAGEFVSDNSNQFKIEVVNDNLNAYKISYCQFGTEKCFNVGRYYDPLTRATRLALSNTPF
```

VFVIKPTDM

利用编码区与非编码区密码子选用频率的差异进行编码区的统计学鉴别方法：由于内含子的进化不受约束，而外显子则受到选择压力，因此内含子的序列要比外显子更随机。这是目前各种预测程序中被广泛应用的一种方法，如GCG(Genetic Computer Group 研制，一种通用核酸、蛋白质分析软件包)的TestCode、美波士顿大学 GeneID 和 Baylor Medcine College 的 BCM Gene Finder 等程序均利用了这一方法。具体方法描述可参阅相关程序说明。

CpG岛：CpG岛(CpG island)一词是用来描述哺乳动物基因组DNA中的一部分序列，其特点是胞嘧啶(C)与鸟嘌呤(G)的总和超过 4 种碱基总和的 50%，即每 10 个核苷酸约出现一次双核苷酸序列CG。具有这种特点的序列仅占基因组DNA总量的 10%左右。从已知的DNA序列统计发现，几乎所有的管家基因(House-Keeping gene)及约占 40%的组织特异性基因的 5'末端含有CpG岛，其序列可能包括基因转录的启动子及第一个外显子。因此，在大规模DNA测序计划中，每发现一个CpG岛，则预示可能在此存在基因。另外，AT含量也可以作为编码区的批示指标之一。

三、序列相似性比较法

近年来相似比较算法也被应用于预测可能存在的基因。这一方法之所以可以预测新基因，主要有以下几个原因：

(1)大约已经有 50%的基因有了对应的 EST，已知的蛋白质序列也越来越多；

(2)不少原核生物和酶母的全序列已经测定。研究表明有将近一半的脊椎动物基因可以通过 BLAST 在酶母、细菌和线虫的序列数据库中找到相似性相当高的序列；

(3)大多数EST 都采用每个克隆分别从 5'和 3'测序，克服了早期EST只代表 3'外显子的缺点。

许多基因预测的程序都已经整合了同源比较算法。

下面举例说明如何通过人类 EST 数据库搜索和拼接与已知基因高度同源的人类新基因：

①以已知基因 cDNA 序列对 EST 数据库进行 BLAST 分析，找出与已知基因 cDNA 序列高度相似的 EST；

②用 Seqlab 的 Fragment Assembly 软件构建重叠群，并找出重叠群的一致(consensus)序列；

③比较各重叠群的一致序列与已知基因关系(图 4.14)。通常有两种情况，一是 EST 足够多，可形成一个覆盖全长的重叠群，以此拼接基因全长序列；另一情况则是，EST 形成几个重叠群，所以可以拼接基因的几段序列。

④对编码区蛋白质序列进行比较，并与已知基因蛋白质的功能域(domain)进行比较分析，推测新基因的功能。

⑤用新基因序列或 EST 序列对 STS 数据库进行 BLAST 分析，如果某一 EST(非重复序列)与某一 STS 有重叠，那么，STS 的位置即确定了新基因的定位。

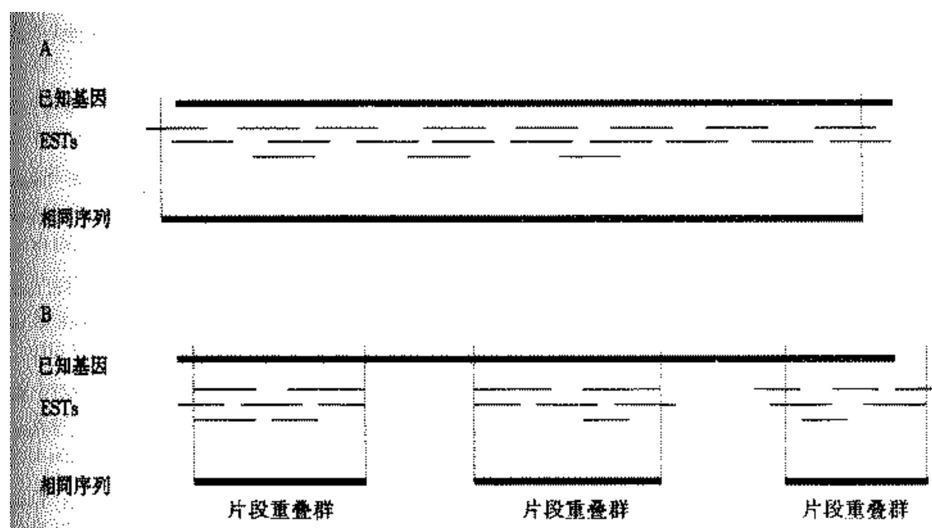


图 4.14 应用以知基因对 EST 数据库进行同源性比较构建的两种 EST 重叠群情况（贺林，2000）

四、隐马尔可夫模型(HMM)

改进目前数据库搜索技术的灵敏性和速度的一条可行办法，是通过蛋白质家族的多序列列线(mutiple alingment)建立一致序列(consensus sequence)。与两条序列的列线比对不同，一致序列可揭示更多的信息，如家族内保守程度不一的残基位置，残基插入和缺失的可能性等等。一致序列的所有表述形式，例profile、模块(block)等都可视为隐马尔可夫链(Hidden Markov Model,HMM)的特例。

HMM 是最近几十年发展起来的时间序列模型，已在语音识别(speech recognition)、离子通道记录、最佳特征识别等方面被应用。HMM 也被较早地应用于生物信息学上的一些问题，如 DNA 编码区、蛋白质超级家族(supper family)的构模等。但是，直至上世纪 90 年代中叶，HMM 才与机器学习技术结合，被系统地应用于整个蛋白质家族和 DNA 区段的建模、列线和分析。HMM 与神经网络、随机模型(stochastic grammar)和贝叶斯网络(Bayesian networks)关系极其密切，或者可视为它们的一个特例。HMM 将 DNA 序列的形成看作一个随机过程，编码和非编码的 DNA 序列在核苷酸选用频率上有所不同而对应于不同的马尔可夫模型。由于这些马尔可夫模型的统计规律是未知的，而 HMM 能够自动寻找出其隐藏的统计规律，因而被称为隐马尔可夫模型。对于处理复杂的 DNA 序列，HMM 需要学习不同 DNA 序列结构的信息。

初阶(first order)或称为 0 阶离散HMM是一种时间序列随机通用模型，由有限的状态集S、离散字符表A、转换(transition)概率矩阵T=(t_{ji})和散发(emission)概率矩阵E=(e_{ix})定义。字符散发，系统由一种状态随机地向另一种状态进化。假设系统处于状态i，它存在t_{ji}概率转变为状态j，而字符x散发的概率为e_{ix}。因此，对于HMM来说，系统的每一个状态只与 2 个不同的骰子(dice)节点有关：散发节点和转换节点。0 阶马尔可夫链假设散发和转换仅由现状态决定，而与过去的状态无关。而字符的散发只有模型系统本身可以识别，即所谓“ 隐藏 ”

(hidden)。

图 4.15 给出了一个非常简单的HMM例子。例子中，最后观察到的序列为ATCCTTTTTTTCA。我们可以想象有 2 个“DNA节点”(DNA dice)：第一个节点的散发概率向量为(e_{1A}=0.25,e_{1C}=0.25,e_{1G}=0.25,e_{1T}=0.25)，第二个节点的散发概率向量为(e_{2A}=0.1,e_{2C}=0.1,e_{2G}=0.1,e_{2T}=0.7)。而转换概率如图所示。

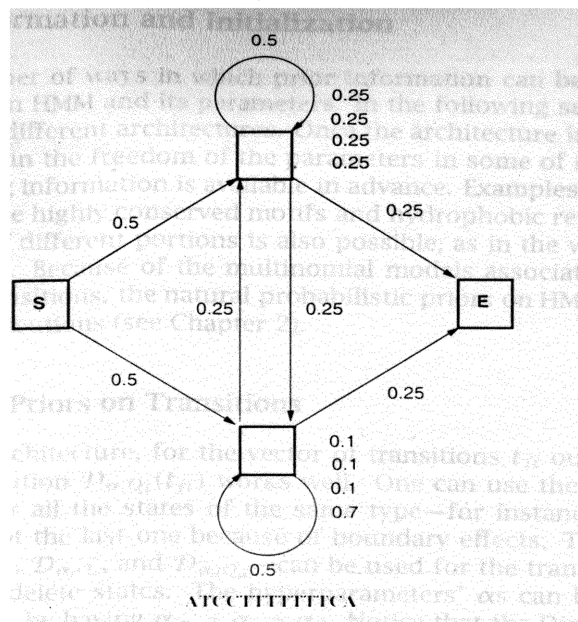


图 4.15 一个简单的 HMM 例子。例子中除了开始（S）和结束（E）两个状态外，还有两个中间状态（Baldi and brunak, 1998）。

对于生物序列而言，HMM 的字符当然是 20 个字母的氨基酸或 4 个字母的核苷酸。但依据不同的问题，其它的一些字符也可使用，如 64 个字母的三联体字母，3 个字母(,β,coil)的二级结构等。当然，HMM 模型并非如上所举的仅有 2 个节点例子那么简单。图 4.16 给出了一个最基本和被广泛应用的“左－右”(left－right)结构模型——标准线性结构模型。所谓“左－右”结构是指该结构中不存在从一种状况回复到已有状况的情况。对于 HMM 模型，一个蛋白质家族如同语音识别中一个词的不同语调。

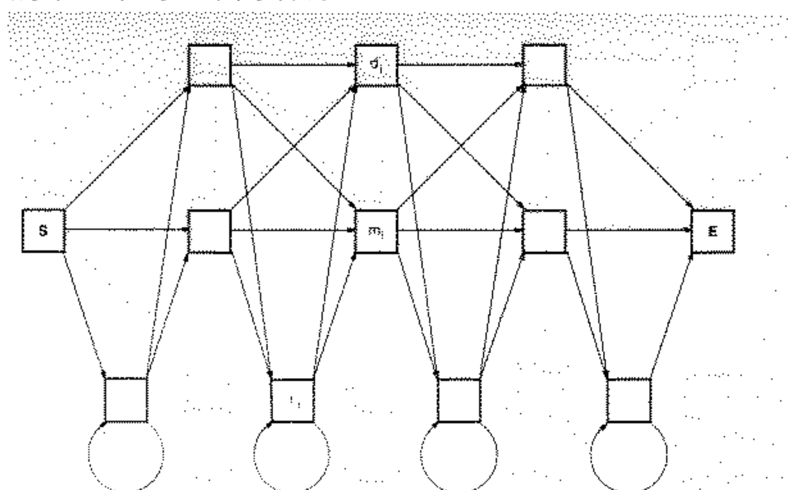


图 4.16 HMM的标致结构。S和E表示开始和结束状态，d_i、m_i和i_i分别表示缺失、

维持和插入状态（Baldi and brunak, 1998）。

一旦一个蛋白质家族成功地构建了 HMM 模型,则该模型可以用于多个领域：①多序列列线；②数据库序列数据的挖掘和分类；③结构分析和模式查找。

五、神经网络

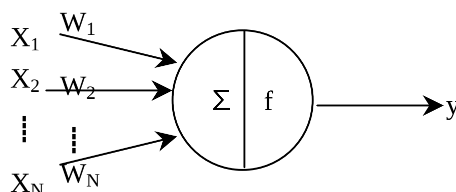
1、神经网络的基本原理

神经网络(NN)属于信息科学理论范畴，它是随着信息科学的开创而发展起来的，目前发起的“神经计算机”革命对计算机产业产生了空前的推动作用。

神经网络是由大量的简单处理单元(即神经元)构成的非线性动力学系统，它具有的学习算法能使其对事物和环境具有很强的自学习、自适应和自组织能力。它能解决常规信息处理方法难以解决或无法解决的问题，尤其是那些属于思维(形象思维)和推理方面的问题。

在人工神经网络中，神经元常被称为“处理单元”，有时从网络的观点出发又把它称为“节点”。人工神经元是生物神经元的一种近似，在功能上讲它只是一阶逼所。它仅仅近似地模拟了生物神经元的三个过程。

处理单元的结构：



（1）输入与输出

（2）加权系数

（3）神经元函数：如常用的活化函数 S 型(Sigmoid)函数

目前应用的一些神经模型包括：感知器(perceptron)模型、反向传播网络(backpropagation network, BP 或 BPN)模型、自组织特征映谢模型(self-organizing feature map, SOFM)、回归网络(recurrent network)模型(Hopfield 提出)、混合网络和混合系统模型。其中混合系统模型是指把神经网络与常规信号处理系统模型结合起来，以便分别取其所长，目前在语音信号处理中将神经网络与隐马尔柯夫模型(HMM)结合起来进行语音识别即是一例，同时在生物信息学上除了应用 NN(BPN)外，也将此混合系统模型加以应用。

神经网络的学习规则：神经网络中的神经元是一个具有相当自适应能力的处理单元，它所连接的权可以根据一定的规则来调整。比较流行的几种规则：

（1）Hebb 规则

（2）感知器学习规则

（3）δ学习规则

（4）Widrow-Hoff 学习规则

（5）相关学习规则

（6）胜者取全学习规则

2、BPN 神经网络

以下重点介绍 BPN 神经网络。

BPN网络由输入层、输出层以及若干隐层节点互连而成的一种多层网，它的输入和输出是在[0，1]或[-1，+1]区间连续取值，每个处理单元对输入的加权和y_i加以非线性处理，得到其活性输出。最常用的非线性函数为Sigmoid函数：

$$f_{(yi)} = \begin{cases} \dfrac{1}{1+e^{-y}} & (0,1) \quad (f_{(x)} = \dfrac{1}{1+e^{-x}}) \\ \dfrac{1-e^{-y}}{1+e^{-y}} & (-1,+1) \end{cases}$$

BPN为前馈网络，对其训练所采用的算法是反向传播法，这是一种有导师学习方法。它利用了均方误差和梯度下降法来实现对网络连接权的修正。对网络权值修正的目标是使网络实际输出与规定输出之间的均方误差(mean squared error，MSE)最小。对于一个处理单元的情况下，如果网络有K个训练样本{E^k}，对应的正确输出为{C^k}，网络的权为W，则用ε表示MSE为：

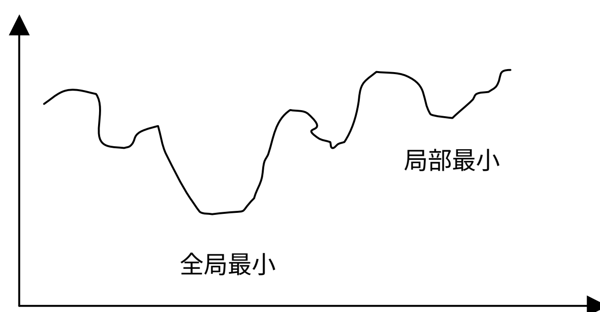
$$\varepsilon = \frac{1}{k}\sum_{k=1}^{k}(w \cdot E^k - C^k)^2$$

它可以看成是权的函数ε(w)，则我们可以按下式来修正权值：

$$W^* = W - \rho \triangle \varepsilon(W)$$

其中，ρ是一个大于零的小数，它规定了修改的步幅。

梯度下降法的基本思想：首先设置权 W 的一组初值，然后，连接计算均方误差相对于权的梯度，并按上式一小步小步地修正权值，当满足一定的准则时（比如 MSE 进入到下限的某一范围时）即停止。这时称为算法收敛。对于梯度下降算法来说，最大的问题是不能保证收敛到全局最优。



更新权值公式可进一步分解为：

$$W_{ij}^{*}=Wij+\rho\delta_i X_j,\quad \delta i=-\frac{\partial\varepsilon}{\partial y_i}$$
$$\downarrow$$
$$W_{ij}^{*}=W_{ij}+\alpha\Delta W_{ij}+\rho\delta_i X_j,\quad \Delta W_{ij}=W_{ij}^{*}-W_{ij}$$

一般ρ选 0.1 以下，α选 0.9，初始权值可在(-2/q，2/q)之间选择(q为一个神经元的输入数)。δ_i为求误差梯变过程的一个中介量。

新的梯度法：共轭梯度(conjugate gradient)法和准牛顿(quasi-Newton)法。

克服局部极小问题：可随机换一组初始权值重新训练一次。

3．神经网络的应用

神经网络技术应用于生物序列分析领域已有较长历史。1982 年利用氨基酸序列，神经网络便被用于预测核糖体结合位点。但是神经网络在本领域的真正应用，是在 1986 年多层神经网反向传播(即 BPN)学习算法被广泛应用后才开始的，特别是 1988 年该技术被应用于蛋白质二级结构预测后，该技术已在多方面取得了很好的应用效果。一些程序已采用了神经网络方法，例如比较有名的 GRAIL 程序(Gene Recognition and Analysis Internet Link)(由美 Oak Ridge 国家实验室研制，http://avalon.epm.ornl.gov/Grail-bin/EmptyGrailForm/)。GRAIL 采用了神经网络技术，具有根据范例而“学习”的功能。向 GRAIL 程序提供了一组已知序列，其中序列的外显子和内含子的位置都已通过实验确定；神经网络根据设计运行，确定特定类型的简化特性，当一条待测序列输入后，网络可以利用已建立的序列与特性之间的关系，找出待测序列的外显子等。

神经网络主要应用于以下几个方面：

①序列编码分析；
②蛋白质二级结构预测；
③单肽及其切割位点预测；
④遗传密码的结构和起源分析；
⑤真核生物基因寻找和内含子剪接位点预测。

六、RNA 二级结构预测

尽管现有一些 RNA 折叠程序可以预测 RNA 二级结构，但这类分析仍然是一

门艺术。RNA 折叠有助于找出 RNA 分子中可能的稳定茎区，但对给定的 RNA 分子来说，这一结果的生物学意义究竟有多大，还是一个未知数。即使有此局限性，二级结构的预测还是有助于找出 mRNA 控制区以及 RNA 分子中可能形成稳定折叠结构的区段。

预测二级结构的最大难题是对三级结构中既有的相互作用进行模型处理，然后将此处理结果回归成一级结构要素，以用于折叠结构的预测。诚然，现有的 RNA 折叠程序并未考虑核酸分子中可能的三级结构。这些程序只能定出有限数目的二维结构的能学参数，由此推测的二维最稳定结构，可能与三维最稳定结构相去甚远，因为三维亿个结构里的环区可以与环区相互作用，螺旋区可以堆积，还会出现各种的非 Watson-Crick 碱基对结构。

目前已有一些比较有名的预测程序，例如 MFOLD [M 代表多(multi)，从早期的 RNAFold 程序或 GCG 软件包的 FOLD 程序扩充而成]，由加拿大国家研究基金会的 Michael Zuker 设计。除对碱基配对的标准能学进行分析外，MFOLD 还考虑到了碱基堆积的能量及单碱基统计的熵。这一程序的 VMS、VNIX、DOS 和 Macintosh 版本可以从许多软件组合中找到。尽管 MFOLD 的输出是文本形式的(图 4.17A)，但有几个程序可以将预测结构转化为图示形成(例如由 Don Gillbert 设计的 Loop Viewer，见图 4.17B)。

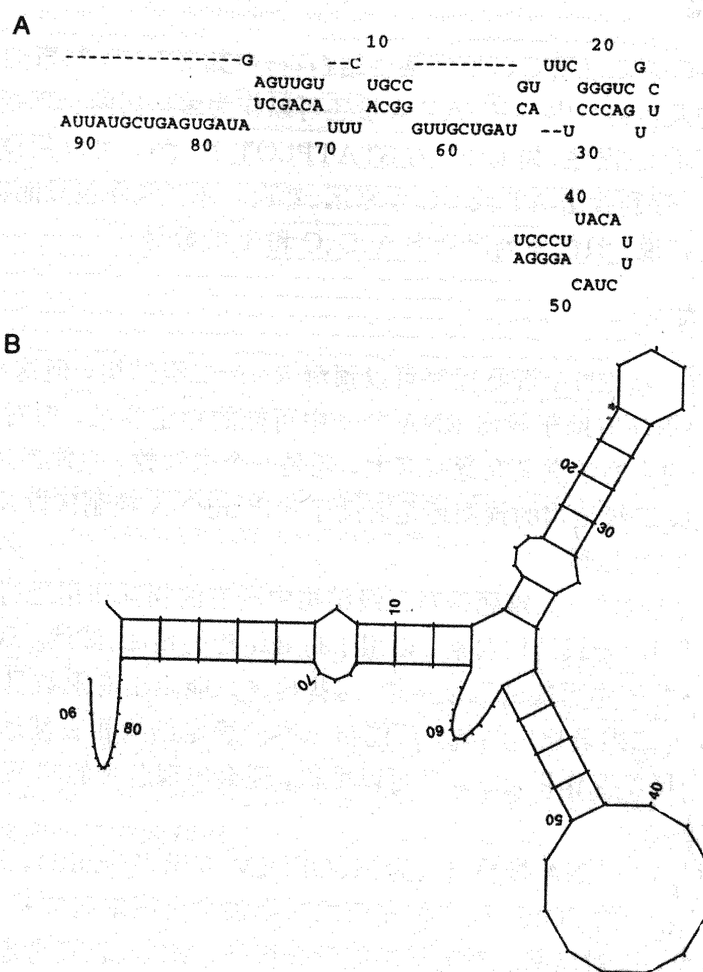


图 4.17 RNA 二级结构的文本输出结果（A）和图形显示(B)。分别由 GCG 的 FOLD 和 Squiggles 程序生成。

第三节 基因组分析

一、基因组分析：生物信息学发展的“史记”

自从1995年第一个可以独立生存的生物被基因组测序以来（Fleischmann et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae*. *Science*. 1995，269:496-512），每年在*NATURE*和*SCIENCE*杂志上都会发表一些重要生物基因组测序完成后的分析文章。这些大文章（Article）中对基因组的分析可谓登峰造极，往往包括了当时想得到的和可以做得到的序列分析手段，它们代表着当时生物信息学发展的最新高度。可以说，这些文章是生物信息学发展史的另类记录。

以下列出了一些重要基因组分析文章，感兴趣的读者不妨对他们的分析内容或方法做些比较：

1977 First biology: Phage ϕX174 (5.386kb)

Sanger F, Air G M, Barrell B G, *et al.* Nucleotide sequence of bacteriophage phi X174 DNA. *Nature*, 1977, 265:687-695

1982 Phage lambda genome

Sanger F, Coulson AR, Hong GF, Hill DF, Petersen GB. Nucleotide sequence of bacteriophage lambda DNA. *J Mol Biol.* 1982, Dec 25;162(4):729-73

1983 Phage T7 genome (39.937kb)

Dunn,J.J. and Studier,F.W. Complete nucleotide sequence of bacteriophage T7 DNA and the locations of T7 genetic elements. *J. Mol. Biol*. 1983, 166 (4), 477-535

1995 First bacterial genomes (1.8 Mb)

Fleischmann et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*. 1995 Jul 28;269(5223):496-512

1996 Yeast genome

Genome sequence of the yeast *S. cerevisiae* Overview of the yeast genome. H. W. MEWES *et al*. *Nature* 387, suppl. 7-8 (29 May 1997)

1997 *E. coli* genome

The Complete Genome Sequence of *Escherichia coli* K-12. Frederick R. Blattner, et al. *Science*, Volume 277, Number 5331, Issue of 5 Sep 1997, pp. 1453-1462.

1998 Worm (multicellular) genome

Genome Sequence of the Nematode *C. elegans*: A Platform for Investigating Biology. The *C. elegans* Sequencing Consortium. *Science* Dec 11 1998: 2012-2018.

1999 Fly genome

The Genome Sequence of *Drosophila melanogaster.* Mark D. Adams, et al. *Science* Mar 24 2000: 2185-2195.

2000 First plant genome: *Arabidopsis thaliana*

Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana.* THE ARABIDOPSIS GENOME INITIATIVE. *Nature* 408, 796-815 (14

December 2000)

2001 Human genome

The Sequence of the Human Genome. J. Craig Venter, et al. *Science* Feb 16 2001: 1304-1351.

Initial sequencing and analysis of the human genome. THE GENOME INTERNATIONAL SEQUENCING CONSORTIUM. *Nature* 409, 860-921 (15 February 2001)

2002 First crop genome: Rice (ssp. *indica* and *japonica*) genomes

A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *indica*). Jun Yu, et al. *Science* Apr 5 2002: 79-92.

A Draft Sequence of the Rice Genome (*Oryza sativa* L. ssp. *japonica*). Stephen A. Goff, et al. *Science* Apr 5 2002: 92-100.

Sequence and analysis of rice chromosome 4. Qi Feng, et al. *Nature* 420, 316 - 320 (21 Nov 2002) Letters to Nature

The genome sequence and structure of rice chromosome 1. Takuji Sasaki, et al. *Nature* 420, 312 - 316 (21 Nov 2002) Letters to Nature

In-Depth View of Structure, Activity, and Evolution of Rice Chromosome 10. The Rice Chromosome 10 Sequencing Consortium. *Science* Jun 6 2003: 1566-1569.

2003 Dog genome

The Dog Genome: Survey Sequencing and Comparative Analysis. Kirkness *et al*. *Science*, Volume 301, Number 5641, Issue of 26 Sep 2003, pp. 1898-1903

2004 Rat genome

Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Rat Genome Sequencing Project Consortium. *Nature* 428, 493-521 (1 Apr 2004)

二、比较基因组学[4]

比较基因组学是基因组学的重要分支,它是随着人类和其它生物基因组的大规模测序发展起来的新科学,现已成为研究生物基因组最重要的策略与手段之一。

与比较解剖学、比较组织学等学科一样，比较基因组学使用是遗传学的重要方法——异同的比较，但该学科的特点是在整个基因组的层次上比较，如基因组的大小、基因数量的多少、特定基因的存在或缺失、基因(或标记序列片段)的位置及排列顺序、特定基因或片段的组织等等。而最重要，也是最体现比较基因组学科特点的是全基因组的核苷酸序列的整体比较。随着世界各国基因组计划的实施，除了人类基因组，许多模式生物基因组测序也已完成或正在进行中，如大肠杆菌、酶母、果蝇、线虫、小鼠、鱼、拟南芥等。同时，美国的“食物基因组计划”，几乎包括了所有重要作物：小麦、玉米、大豆、马铃薯、南瓜、棉花，

[4]本部分内容取自陈竺、杨焕明等人的文章，见：贺林．解码生命—人类基因组计划和后基因组计划，北京：科学出版社，2000

而我国的水稻、家蚕、微生物等基因组计划也在进行中或已完成。这些基因组全序列数据将成为比较基因组的最基本研究对象。

认同所有生物的基因组都有共同的进化史，即进化上的共性是比较基因组学的理论依据，可以说，没有进化上的关系，就没有比较基因组学。进化是基因组比较的最重要主题，所以目前基因组比较的生物信息学方法主要来自系统进化分析的一些方法，例如系统进化树的构建方法等。相关内容请参见第五章。基因组比较急需发展针对整个基因组的专用算法。基因组是一种具有大尺度、巨量特点的研究对象，它有其特有问题，必须有特定的算法才能充分挖掘和利用基因组信息。

以下对基因组学分析中经常涉及的四个最基本概念进行介绍：

1、相似性：

相似性(similarity,有时也用analogy)就是简单比较得出的两者之间的相同程度。相似性本身的含义，并不要求与进化起源是否同一，与亲缘关系的远近、甚至于结构与功能有什么联系。核苷酸与氨基酸序列的测定，使原先“模糊”的描述有了定量的指标——百分比。不同基因组之间、不同基因或不同物种的“同一”基因，都可以用%来表示异同程度。

2、同源性：

同源性(homology)是具有严格定义的进化学词汇：在进化上起源同一。同源性可以用来描述染色体—“同源染色体”、基因—“同源基因”和基因组的一个片断—“同源片断”。

在进化上起源同一的两段核苷酸序列，特别是功能较重要的保守区断或基因，一般表现为相似。迄今有证据表明，同源的基因的的确确在核苷酸(或氨基酸)序列上具有较高程度的相似，这就带来了这两个词的混用。如我们有时把“相似搜索(similarity searching)”说成是“同源搜索(homology searching)”。在比较两段序列时，正常的描述应该是：这两个片断可能同源(或这两个基因有可能为同源基因)，因为它们的核苷酸(或氨基酸)的相似程度为 80%。“80%的同源”的说法是不正确的(还有 20%的不同源?)，也是不符合事实与定义的。

相似性与同源性是两个不同的概念，相互之间并没有直接的等同关系。相似的不一定同源，因为在进化的过程中，来源不同的基因或序列由于不同的独立突变而“趋同”并不罕见；同源一般表现为相似，但同源并不一定比非同源的相似程度要高。我们只是在进化的过程的一个时间点上加以观察。功能相似或相同也不一定必然同源。非同源基因的代谢功能替换已有不少证据，其它表型相似也不一定反映了同源，不同基因的不同突变就有可能产生“表型模拟”。

而同源又有两种不同的情况即垂直方向的(orthology)与水平方向的(paralogy)。

3、直系同源：

直系同源(orthology)是比较基因组学中最重要的定义。直系同源的定义是：

(1)在进化上起源于一个始祖基因并垂直传递(vertical descent)的同源基因；

(2)分布于两种或两种以上物种的基因组；

(3)功能高度保守乃至于近乎相同，甚至于其在近缘物种可以相互替换；

(4)结构相似；

(5)组织特异性与亚细胞分布相似。

在这些条件中，垂直传递和功能相同是最重要的。如多种抗药性基因，在细菌、果蝇、河豚鱼、小鼠、人类的基因组中都存在，其结构相似，功能都与多种药物的抗性有关。直系同源基因的鉴定是比较基因组的研究线索和内容，直系同源的存在是基因组进化的重要证据，因此对直系同源的定义与条件的掌握甚为严格。鉴定直系同源的实际操作标准(practical criteria)为：

如基因组Ⅰ中的A基因与基因组Ⅱ中的A´基因被认为是直系同源，则要求：

(1)A´的产物比任何在基因组Ⅱ中所发现的其它基因产物都更相似于A产物；

(2)A´与A的相似程度比在任何一个亲缘关系较远的基因组中的任一基因都要高；

(3)A编码的蛋白与A´编码的蛋白要从头到尾都能并排比较，即含有相似以至于相同的模序(motif)。

3、旁系同源：

旁系同源(paralogy)基因是指同一基因组(或同系物种的基因组)中，由于始祖基因的加倍而横向(horizontal)产生的几个同源基因。

直系与旁系的共性是同源，都源于各自的始祖基因。其区别在于：在进化起源上，直系同源是强调在不同基因组中的垂直传递，旁系同源则是在同一基因组中的横向加倍；在功能上，直系同源要求功能高度相似，而旁系同源在定义上对功能上没有严格要求，可能相似，但也可能并不相似(尽管结构上具一定程度的相似)，甚至于没有功能(如基因家族中的假基因)。旁系同源的功能变异可能是横向加倍后的重排变异或进化上获得了另一功能，其功能相似也许只是机械式的相关(mechanistically related)，或非直系同源基因取代新产生的非亲缘或远缘蛋白在不同物种具有相似的功能。在真细菌与古细菌的基因组中，30%～50%的基因属旁系同源，在真核基因组的比例更高(Koonin EV and Galperin MY,1997)。

相似与同源，直系与旁系需要在定义上加以明确，但实际应用中很难截然分开。与别的常用术语也很难明确界定。但基因家族或多基因家族(gene family, multigene family)的原来的定义较侧重于结构，因而一个直系基因可以与几个旁系基因同属于一个基因家族。在这一定义上，旁系同源可以说是一个基因家族中的其他成员(Huynen et al, 1997)。

随着不同物种全基因组序列的阐明，上述概念愈见重要并更明确。从已知的 7 个物种的全基因组序列比较，如所有的保守基因都据同源关系而加以分类(Tatusov RL et al.,1997)，可归纳出 720 个直系同源簇(clusters of orthologous groups,COG)，每一 COG 由一个直系同源蛋白或存在于至少 3 个种系(lineage)的直系的旁系同源组(orthologous sets of paralogs)组成。而基因家族又因大批基因及产物序列而赋予新的内容，这对于扩大对生物过程的认识与操作基因的能力有很大的意义(Henikoff et al.,1997)。

第四节 基因组分析列举：水稻基因组分析

本节将结合我们近年来的一些研究结果，重点对第一个被基因组测序的作物——水稻的基因组研究和分析结果进行介绍。

水稻是第一个被全基因组测序的作物。亚洲栽培稻（*Oryza sativa*）共有2个亚种（籼稻和粳稻），其中一个粳稻品种"日本晴"分别通过全基因组鸟枪法（Goff et al, 2002）和逐步克隆方法(Sasaki et al, 2002; Feng et al, 2002; The Rice Chromosome 10 Sequencing Consortium, 2003; The Rice Genome Sequencing Project, 2005)测序，另一个籼稻品种"9311"通过全基因组鸟枪法测序(Yu et al, 2002; Yu et al, 2005)。除了核基因组外，水稻的叶绿体基因组序列早在15年前就已测序完成(Hiratsuka et al, 1989)，同时，其线粒体基因组最近也被测序完成（Notsu et al. 2002）。

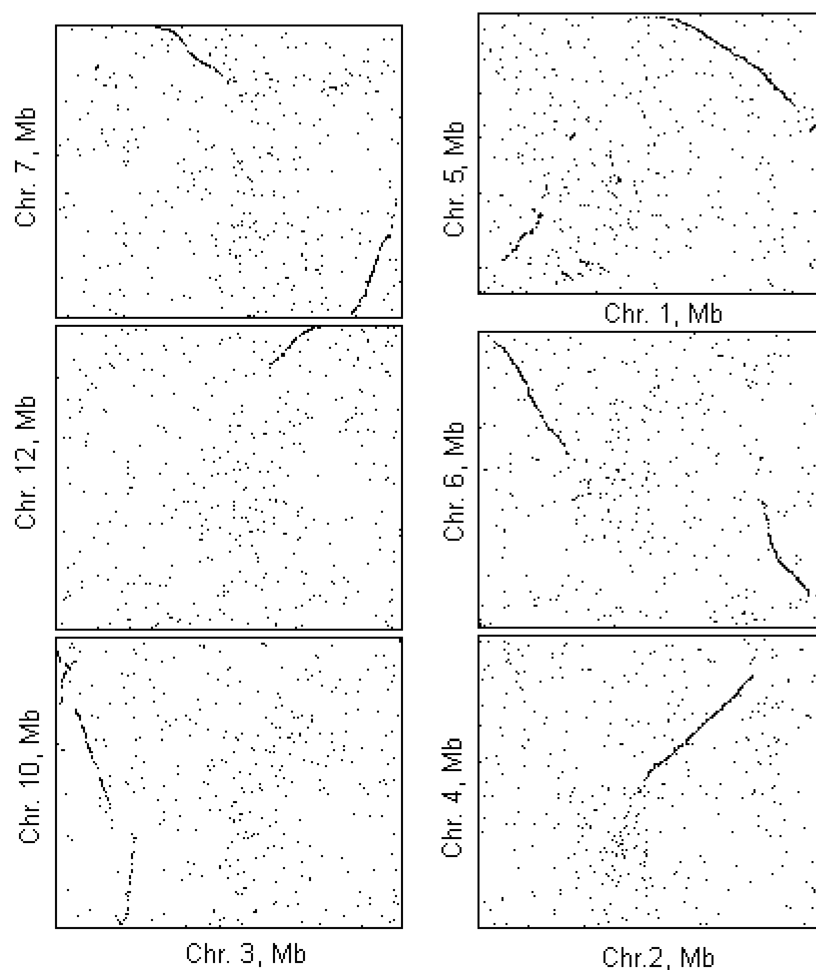
在获得基因组序列后，一项艰巨的研究任务是如何从巨量的水稻基因组序列中挖掘出潜藏的遗传事件、进化机制等重要生物信息。为此本文结合我们自身的一些研究工作，重点介绍了近年来在水稻基因组序列分析中获得的几项最新的研究结果。

1 现代的二倍体，古老的多倍体

2004年水稻基因组研究的一个重要进展，是获得清晰的证据表明水稻基因组曾发生过全基因组倍增。Paterson等(2004)、Guyot 等(2004)和我们(Fan et al, 2004;Zhang et al, 2005a)的研究结果也一致表明，在禾本科作物分化前发生过一次全基因组倍增（whole-genome duplication）。早在2002年，根据最初的

水稻基因组草图序列,Goff 等（Goff et al, 2002）利用同义替换率分布方法（K_S-based age distribution）提出水稻基因组可能发生过一次全基因组倍增。而在此之前，利用分子标记、DNA 重复元件等方法对水稻部分染色体区段的研究，也提出水稻基因组的一些染色体间可能发生过片段倍增（block or segmental duplication）。2003 年两篇重要文章相继发表，对水稻基因组起源和倍增事件做出了初步分析和有益探索(Paterson et al, 2003; Vandepoele et al, 2003)。随着水稻基因组序列数据的增加，特别是美国基因组研究院（TIGR）利用逐步克隆(clone by clone)测序的数据首次拼成 12 条水稻染色体序列，利用 TIGR 的数据和基因相似性矩阵方法（GHM, gene homology matrix）, 检测到大量染色体间的倍增片段，这些倍增片段几乎覆盖了水稻全基因组（图 1,图中包括水稻第 2 号染色体与第 4 和 6 号染色体、第 3 号染色体与第 7、10 和 12 号染色体和第 1 与 5 号染色体间的间的倍增片段。另外第 8 与 9 号染色体、第 11 与 12 号染色体间的倍增片段未列出）。这是全基因组倍增的有力证据。根据倍增片段上同源基因的分子进化分析，全基因组倍增大致发生在 7000 万年前，在禾本科作物分化前(Paterson et al, 2004)。我们在 2004 年初利用 TIGR 的第一版水稻基因组数据（osa1, Version 1）和 GHM 方法就已发现了这一水稻基因组倍增的证据并投稿 (论文摘要已递交上海-合肥举行的系统与进化研讨会,(Fan et al, 2004)。但就在 6 月低-7 月初，Paterson 等(2004)和 Guyot 等(2004)的文章相继发表。后我们利用 TIGR 更新的数据（osa1, Version 2）对水稻染色体间倍增片段进行了更新，并以此为基础，利用同义替换率分布方法检测到另一次更古老的（单双子叶植物分化前）基因组倍增事件(Zhang et al, 2005)。该研究的最新进展是中科院北京基因组研究所（华大）刚刚发表的水稻基因组精细图分析结果也同样证实

了水稻基因组的倍增(Yu et al, 2005)，同时，另外一个独立的课题组最近也获得了同样的结论(Wang et al, 2005)。



引自 Zhang 等（2005)

图 1 部分水稻基因组倍增片段

全基因组倍增或整倍体化过程被认为是植物尤其是禾本科作物物种形成和进化过程中非常普遍和重要的事件，50%-70%的开花植物在进化过程中均经历了一次或多次染色体加倍过程(Wendel et al, 2000)。基因组加倍后，再经历所谓的二倍体化过程（diploidization），进化成当代的二倍体物种。大量的复制基因将在二倍体化过程中丢失。整倍体化过程一般可通过同源加倍（autopolyploid）

和异源加倍（allopolyploid）两种方式发生。已测序完成的模式植物拟南芥，经全基因组序列分析发现，至少发生过 3 次全基因组自身复制(Bowers et al, 2003)；玉米被认为在其与高粱分化后发生一次异源加倍过程，即起源于异源四倍体(allotetrapolyploid)。利用同义替换率分布方法检测和最新序列数据库数据，Blanc 和 Wolfe(2004)在很多重要作物中均发现了全基因组倍增的证据。

水稻全基因组倍增片段是迄今为止发现的在动植物基因中最为清晰、完整的基因组倍增的遗迹。拟南芥基因组在更近代的时候也发生过全基因组倍增，但它的倍增片段都比较短且凌乱(Bowers et al, 2003; Simillion et al, 2002)。水稻之所以保存得这么完整可能与水稻基因组相对比较稳定有关(Llic et al, 2003)。

2 最小的核基因组：基因组在扩增还是在缩小？

植物界基因组中 DNA 含量差异很大，它们的差异性与生物的复杂性程度并不完全相关，这种现象称为 C 值悖理。如大麦（*Hordeumvulgare*）、水稻和拟南芥的生物复杂性比较相似，但大麦基因组分别为水稻和拟南芥基因组的 11 倍和 35 倍。众多因素（机制）决定了基因组的膨胀和缩小(Bennetzen et al, 2002)，早在 19 世纪 30 年代，基因复制就被认为是增长遗传物质的首要机制(Betran et al, 2002)。在植物界中，基因数目的增加通常归因于基因复制、DNA 片断或基因组复制。基因组膨胀的最主要因素为基因组的倍增(Wendel et al, 2000; Grover et al, 2004)。而转座因子的扩增则是另一个推动基因组增加的关键因素。在禾本科内，已报道在最近的 1 千万年内大多数基因组的膨胀由 LTR 逆转座因子的扩增所导致(SanMiguel et al, 1996; Ma et al, 2004)。很明显的，这一机制只能导致基因组膨胀(Bennetzen et al, 2000)，而基因组只是这样一味地膨胀进化吗？并

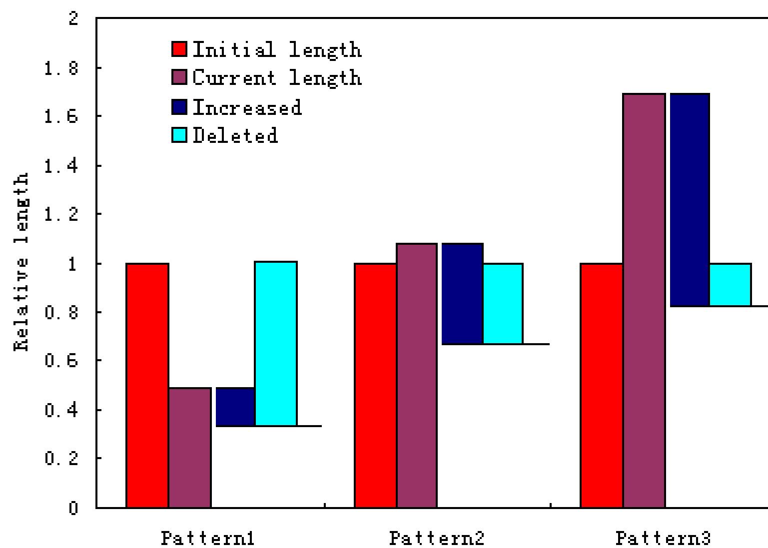
非如此。后来发现了抵制这一膨胀的机制：异常重组(illegitimate recombination)和非同源性重组(unequal homologous recombination)可以减少 LTR 逆转座序列从而抵制基因组的膨胀(Vicient et al, 1999; Shirasu et al, 2000; Ma et al, 2004)。最近已发现水稻和拟南芥基因组中的 LTR 逆转座序列的大量丢失(Ma et al, 2004; Devos et al, 2002)。在最近的 8 百万年里，水稻基因组中至少有 190Mb 的 LTR 逆转座序列被删除(Ma et al, 2004)。利用非洲栽培稻进行的比较基因组学研究表明，亚洲栽培水稻的籼粳稻基因组大小均增加了 2%和 6%(Ma et al, 2004)。但该研究的结论仅是根据约 1Mb 长度的基因组片段（水稻 430Mb 基因组的 0.2%）得出。根据 non-LTR 逆转座研究，Petrov 和他的同事得出非平衡性的少量删除和插入导致昆虫类的基因组缩小(Petrov et al, 2002)。然而，在植物基因组中是否存在同样相似的机制作用于转座因子，或者其它机制导致非重复序列的丢失仍然没有明确的答案。

为了探索基因组大小改变的潜在进化机制，一种较理想的途径是比较基因组间大小差异很大的相近物种。通过比较果蝇（165Mb）和其它两个基因组极大的相近物种 *Laupala* crickets（1910Mb）和 *Podisma* grasshoppers（18150 Mb），发现果蝇 DNA 的大量丢失(Petrov et al, 2002)。最近，通过比较异源多倍体物种棉花（*Gossypium hirsutum*）不同基因组序列片断，探索了该物种基因组大小变化的进化机制(Grover et al, 2004)。

在有花植物中，全基因组倍增是普遍发生的现象，并且被认为在物种进化和分化中起着重要作用(Wendel et al, 2000)。一旦染色体倍增过后，古老多倍体的基因组进化速率加快，在“二倍化”过程中伴随着大量的 DNA 序列的消失以及染色体重排现象(Sasaki et al, 2002)。水稻基因组测序工作的完成(Sasaki et al,

2002; The Rice Chromosome 10 Sequencing Consortium, 2003)为研究水稻基因组的进化史提供了一个前所未有的机会。水稻基因组多倍体的起源已被证实(Paterson et al, 2004; Zhang et al, 2005; Paterson et al, 2003)。多倍化事件估计发生在 70 百万年前，在禾本科分化之前(Paterson et al, 2004)。这一结论是基于许多非重叠的倍增块几乎覆盖了整个基因组这一事实而得出。该研究结果为研究水稻基因和基因组倍增后的二倍体化的进化机制提供了非常好的素材。

当一次复制事件发生，两对应的复制片断或染色体在初始阶段通常应具有同样的大小。但经过长期的进化，其同源的复制片断的大小有可能存在差异。由基因组复制产生的复制块（同源复制块）将经历一次“二倍体化”的剧烈进化过程，伴随着大量的 DNA 序列的丢失。同源复制片断间存在的巨大长度差异为分析基因组膨胀或缩小进化机制提供了有效的途径。在我们的研究中，从水稻全基因组倍增产生的同源复制片断（如来自第 2，3，6，7 和 10 号染色体），由于它们存在着巨大的差异性而被选择为研究对象，用于探索水稻经历多倍化后基因组大小的进化机制。我们的研究表明，在最近 70 百万年里，水稻染色体以不均衡的模式(即染色体长度存在膨胀、平衡和减小 3 种情况)进化着，影响复制片断大小的差异主要由非重复序列的 DNA 丢失引起的，且 LTR 因子的扩增也起着重要作用(Guo et al, 2006)（图 2）。



引自 Guo 等（2006)

图 2 水稻基因组染色体长度变化的三种进化模式

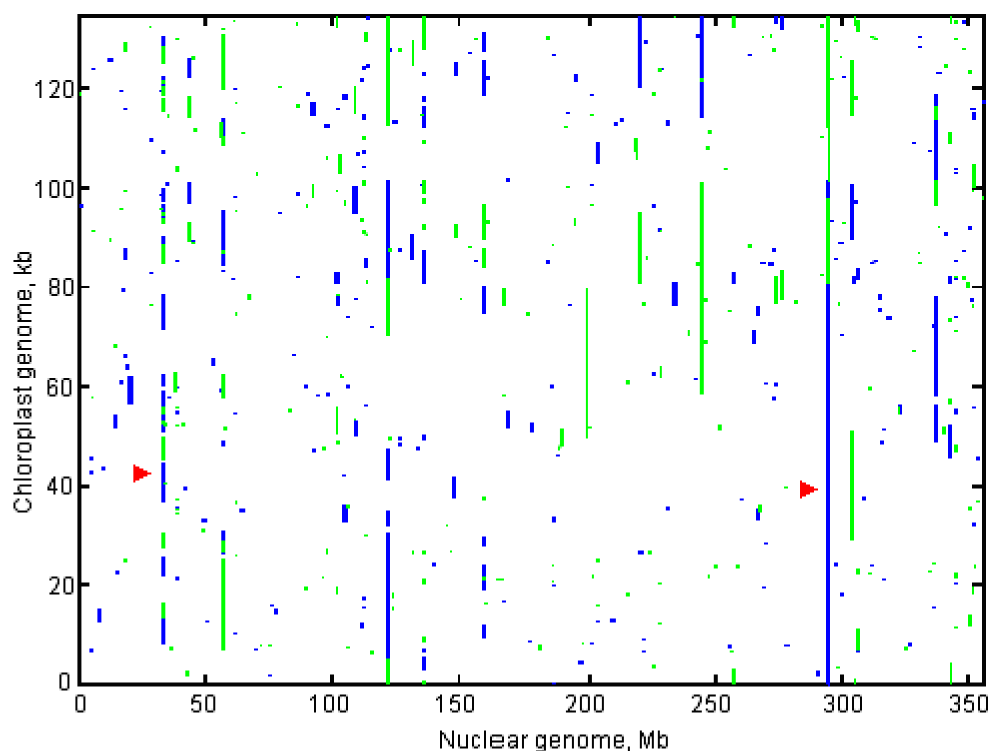
3 籼粳稻分化时间比原来估计的要迟得多

水稻（*Oryza sativa* L.）属于禾本科（Gramineae 或 Poaceae），也是 3 大谷类植物之一，即水稻，小麦（*Triticum aestivum*）和玉米（*Zea mays*）。为人类提供了主要食源。大约在 77 百万年前禾本科从同一祖先分化而来，其两个亚科 Erhartoideae（水稻）和 Panicoideae（玉米和高粱）大约在 50 百万年分开(Gaut et al 2002)。水稻化石的研究可追溯到约 40 百万年前。 22 个水稻物种中已发现 9 个物种为 2 倍体类型（2n = 24）以及由不同重组形成的异源 4 倍体（2n = 48）等。*O.rufigogon* 是栽培稻（*Oryza sativa* L.，AA 基因组）的野生祖先，后被驯化为 *O.sativa*，其驯化时间可能起源于 9 千年前。栽培稻有 2 个主要亚种籼稻和粳稻，基于来自 2 个亚种的 29kb 的同源片断，Bennetzen(2000) 认为它们约在 1 百万年前分开，但是他未给出这一时间估计的详细信息。这一分化时间估计后来在水稻基因组的研究中被广泛引用(Song et al, 2003; Han et al,

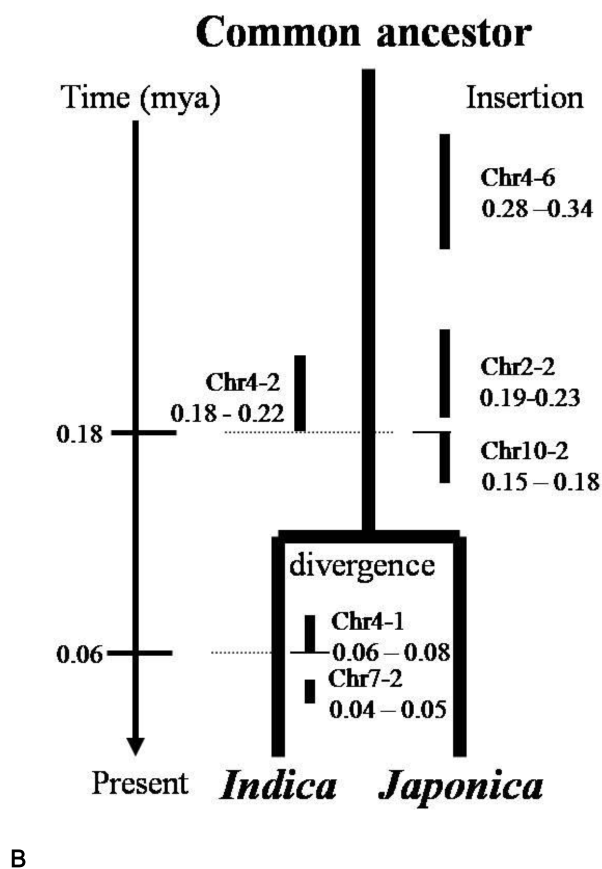
2003)。

水稻基因组测序的工作已基本完成。栽培稻粳稻日本晴通过全基因组鸟枪法(Goff et al, 2002)和利用遗传图和物理图的逐步克隆方法被测定(Sasaki et al 2002; Feng et al, 2002; The Rice Chromosome 10 Sequencing Consortium, 2003)。栽培稻籼稻“9311”通过全基因组鸟枪法被测定(Yu et al, 2002)。除了核基因组，水稻叶绿体基因组早在 15 年前就被测序完成(Hiratsuka et al, 1989)。同样地，玉米和小麦地叶绿体测序工作最近也已完成。

细胞核、叶绿体和线粒体间 DNA 序列的插入很早就被发现(Notsu et al, 2002)。粳稻第 10 号染色体上的 2 个长的叶绿体基因组序列插入已被检测到(The Rice Chromosome 10 Sequencing Consortium, 2003)。同时，籼稻基因组序列中也同样发现大量的叶绿体序列的插入(Shahmuradov et al, 2003)。



A



B

引自 Guo 等（2008a)

图 3 水稻核基因组中叶绿体 DNA 的插入情况 (A)和插入时间估计（B）

植物细胞核和细胞器基因的同义替换率（*ds*）通常被用于进化事件的时间估计(Wolfe et al, 1989)。考虑到叶绿体的一些有利因素，如母系遗传、很少或没有重组等(Sall et al, 2003)，叶绿体 DNA 已被广泛地应用于植物分化时间的估计(Wolfe et al, 1989; Sall et al, 2003; Gaut et al, 2002)。同时，核基因序列也被用于分化时间的研究中，如 Bennetzen 等人的研究(Bennetzen et al, 2000)。非同义替换率(氨基酸改变，*dn*)与同义替换率的(氨基酸不改变，*ds*)的比值(*dn/ds*)也经常被用于分化分析。*dn/ds* 的比值为 1 表示所研究的基因在中性选择(neutral

selection）下进化，小于 0. 25 意味着纯化选择（purifying selection）下进化，当比值大于 1 时则被认为进行正向选择(positive selection)下的进化(Hurst et al, 2002; Swanson et al 2003)。

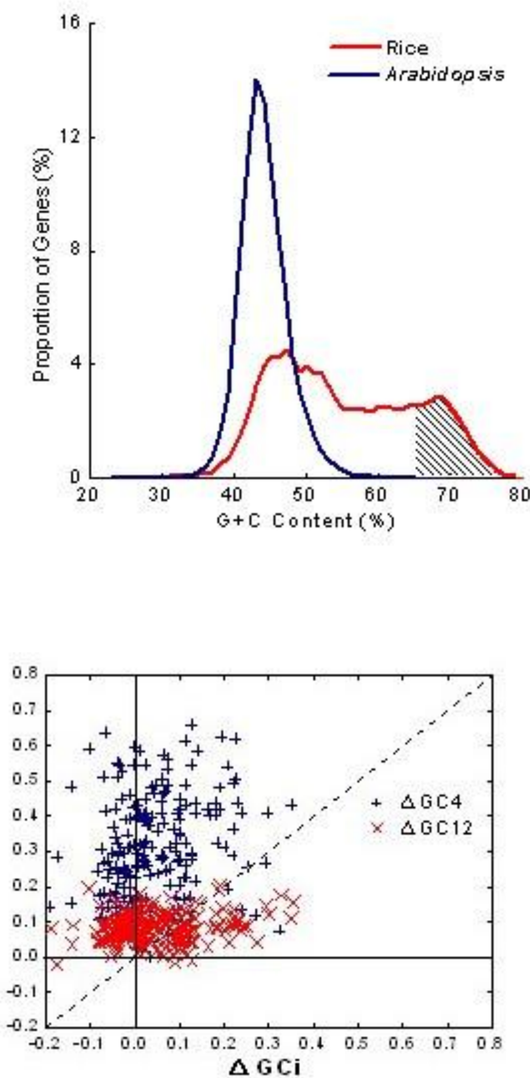
我们利用大片断叶绿体 DNA 的插入（图 3,图中为水稻核基因组序列——12 条染色体依次拼接在一起——与叶绿体基因组联配结果。蓝线表示叶绿体片断顺式插入核基因组,绿线表示叶绿体片断反式插入核基因组）来估计水稻 2 个亚种籼稻和粳稻的分化时间。通过 PCR 扩增和籼稻基因组层次上对叶绿体大片断的搜索 ,确立了籼稻 - 粳稻分化之前叶绿体的最近一次插入并根据同义替换率推断出 2 个亚种分化时间在 6-22 万年之间(Guo et al, 2008a)。该结果与最新一些研究结果基本一致，如利用叶绿体和线粒体基因组序列的研究结果（Tian et al. 2004; Tian et al. 2006)和核基因做出的推断(Ma et al, 2004; Zhu and Ge, 2005; Vitter et al, 2004; Huang et al, 2005)。

4 水稻高 GC 含量基因的进化机制

禾本科基因沿转录方向上 GC（鸟嘌呤＋胞嘧啶）组成上存在负梯度现象最近被发现，而在双子叶植物基因却无此现象(Yu et al, 2002; Wong et al, 2002)。这是一个明显和有趣的现象。但其产生的机制尚无合理的解释。GC 含量作为基因组的一个重要识别标志，已被用于基因组的基本组成的分析，编码序列的进化以及密码子使用偏好性上(Bernardi et al, 2000)。例如，CpG 岛（GC 富含区）被用于真核生物的基因一个路标信息(Ashikawa et al, 2001)。物种间基因平均 GC 含量的变化幅度较大，甚至在同一类物种（如细菌）中也是如此。物种中的这种 GC 含量差异产生的原因尚不清楚。

禾本科内包括了将近 10000 物种，可被分成 700 个属(Gaut et al, 2002)，表现为独立的遗传体系(Bennetzen et al, 1993)。最近的比较基因组学研究表明,所有禾本科植物都追溯到一个共同的"禾本等位基因"(Grass alleles)群体(Freeling et al, 2001)。有报道指出，禾本科有一次 GC 含量提高过程并且在玉米和水稻中可分成两类基因（高 GC 和低 GC）(Carels et al, 2000)。通过考察来自 4 个禾本科物种（水稻，玉米，小麦和大麦）的 25 个基因家族，每个家族成员的基因 GC 含量存在着巨大差异(Zhang et al, 2001)。同时，最近也有报道指出，微卫星分布的一个新特点也沿着基因转录方向呈现梯度变化。对于水稻基因，通常在基因 5' 端能探测到富含 GC 的微卫星，如$(CCG)_n$等(Fujimori et al, 2003)。通过水稻基因组内 CpG 岛的分析，同样也大量出现在表达基因的 5' 端(Ashikawa et al, 2001)。

基于水稻 28000 个全长 cDNA（来源于实验）和基因组序列以及其它物种的类似数据,我们详细研究了禾本科以及其它物种的基因 GC 含量和梯度(Guo et al, 2007)。根据水稻转录组 GC 含量的分布，我们得出了水稻 GC 梯度变化模式和明显的两组基因类型（图 4）。我们发现水稻编码基因由于受到选择效应的影响，密码子使用上存在偏向 GC 的倾向，导致了基因 GC 含量的增加。至少部分水稻基因受到这种机制的影响(Guo et al, 2007)。



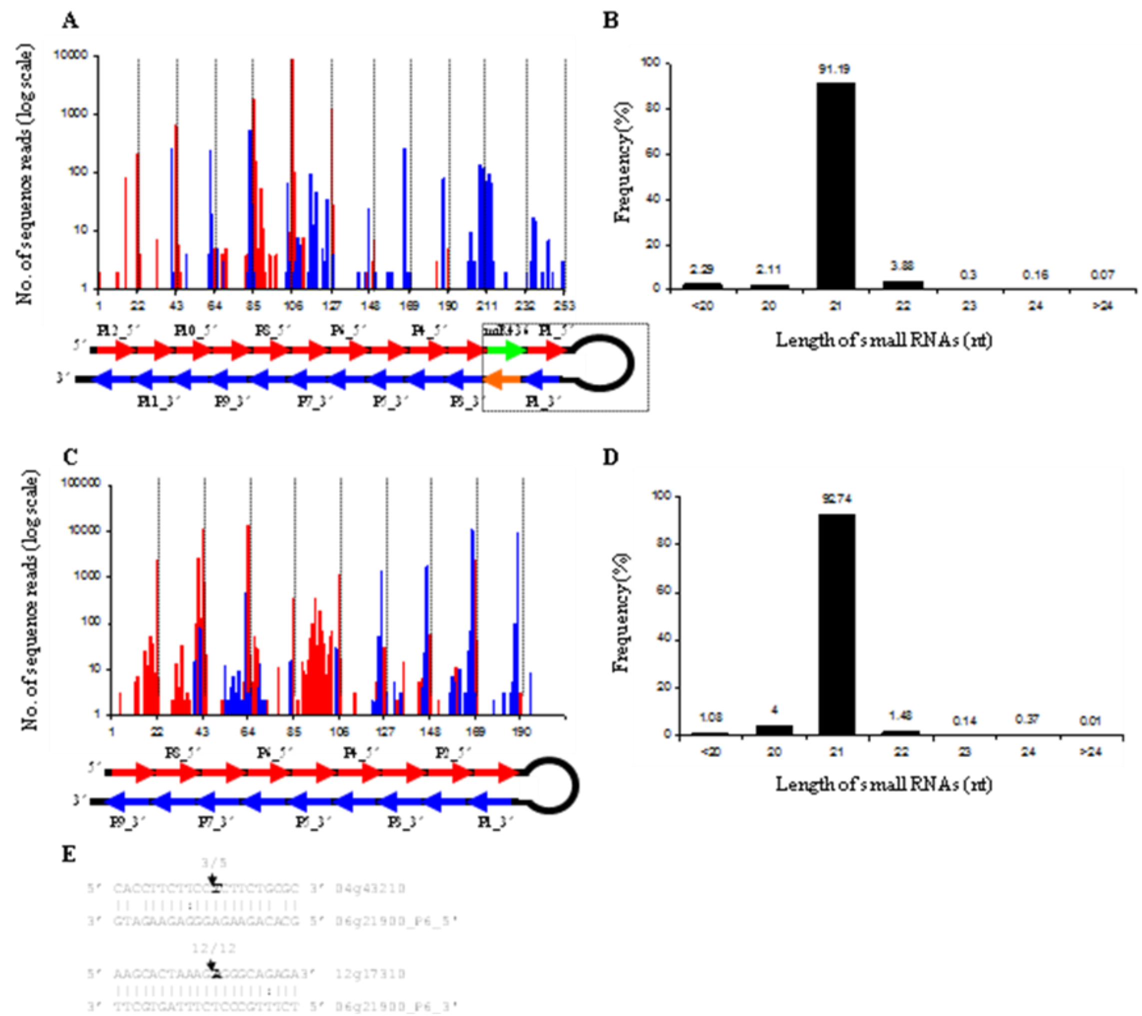
引自 Guo 等（2007)

图 4 水稻和拟南芥基因组基因的 GC 含量分布

5 水稻小 RNA 可能是驯化和育种选择的靶基因

内源性非蛋白质编码的小 RNA(12－24nt)在植物基因转录与后转录水平中起着重要的调节作用。根据小 RNA 的合成机制和功能的不同，可以将其分成两大类，一类是 microRNA(miRNA)，一类是小干扰 RNA(siRNA)。miRNA 是由具

有发夹结构的的初级转录本经过核酸内切酶 DCL1 加工后生成，而小干扰 RNA 则是通过核酸内切酶 DCL2, DCL3 和 DCL4 对双链 RNA 前体进行加工形成的 (Vazquez 2006)。目前在拟南芥、水稻等植物中已经鉴定出了一些小干扰 RNA 位点，包括 ta-siRNAs (trans acting siRNAs)，nat-siRNAs (natural antisense transcript-derived siRNAs)和 ra-siRNAs(repeat-associated siRNAs)，长茎环结构的 miRNA-like 位点（miRNA-like long hairpin）和 nat-miRNA（natural antisense miRNA)。我们鉴定发现了几十个新 miRNA 和一些新类型 siRNA(Zhu et al. 2008）。在水稻中至今已鉴定出 344 个 miRNA (miRBase, http://microrna.sanger.ac.uk/sequences/,Release 12.0)，一个 ta-siRNA 家族（TAS3）, 两个长茎环结构的 miRNA-like 位点和一个 mirtron (Zhu et al. 2008)。

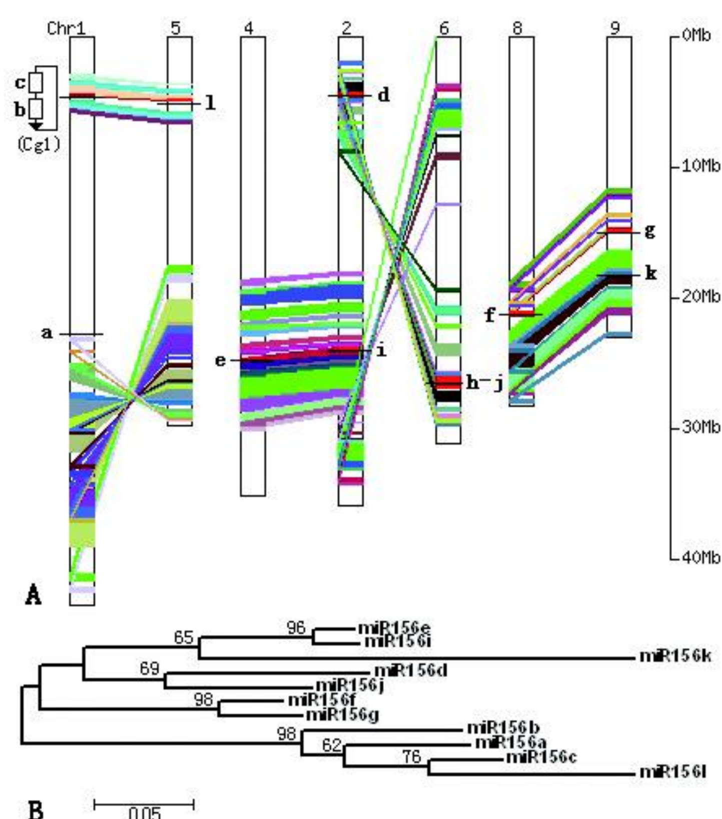


引自 Zhu 等（2008）

图 5 两个长茎环结构的 miRNA-like 位点

遗传学方面近几年的一个重要的研究进展是在动植物基因组中发现了大量小 RNA 等非蛋白质编码基因，这些小基因（一般 100-200bp）在生理生化等代谢过程中起到重要作用。由此产生一个有待回答的问题：这些基因位点在我们人类进行作物驯化和育种过程中是否同样受到选择？我们目前在研究作物骨干亲本遗传成因中是否和如何考虑这些基因对骨干亲本形成的影响？目前发现的人

工选择（育种）的基因位点主要编码转录调节因子和其他蛋白质编码基因，我们的研究发现非蛋白质编码基因在人工驯化过程中同样受到人工选择效应的影响。我们利用水稻为模式作物，发现小 RNA 之一，microRNA 基因 *MIR156b/c* 基因位点可能受到强烈的自然和人工选择效应的影响，说明人工选择的对象除了转录因子及其下游基因外，还可能针对转录因子调控（上游）基因（Wang et al, 2007）。

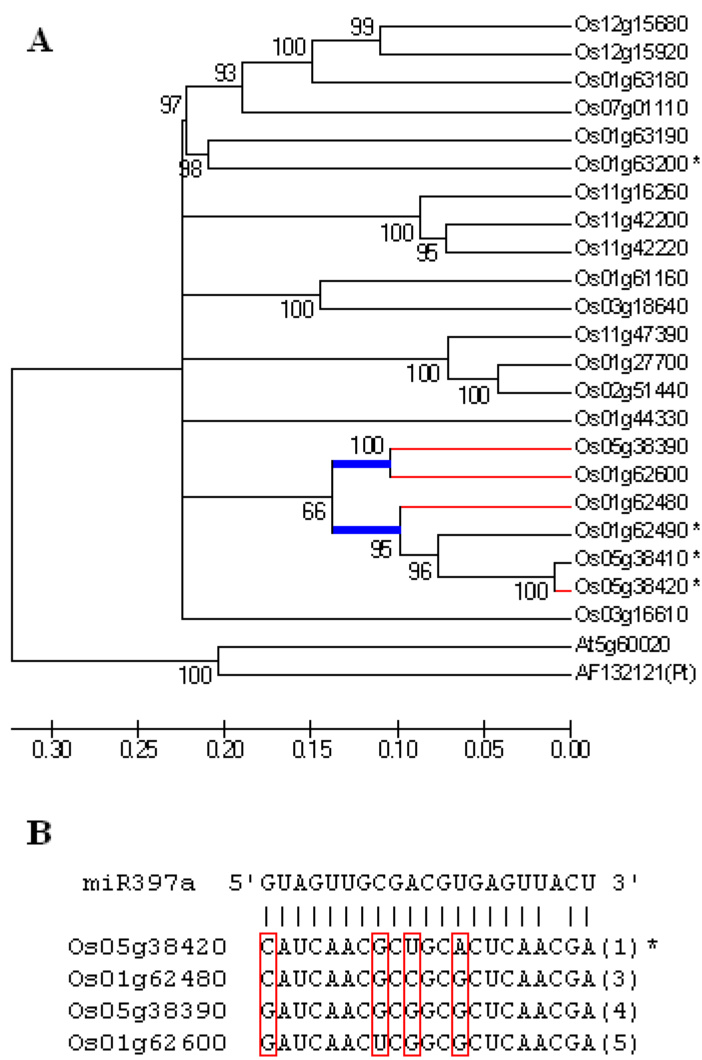


引自 Wang 等（2007）

图 6 水稻 miR156 家族在基因组上的分布和系统进化关系

通过水稻 miRNA 及其靶基因结合位点序列变异的调查和直系同源基因（Paralogs）分析，发现水稻 miRNA 基因在不断地捕获新的结合位点（靶基因），

同时也不断丢失对靶基因的调控功能（Guo et al, 2008b）。这种动态的进化过程主要通过 miRNA 序列突变来实现，同时插入和删除也发挥一定作用。图 7 展示了水稻 miR397 靶基因在全基因组前后的突变进化情况，有些靶基因位点由于序列突变而脱离了 miR397 的绑定和调控。

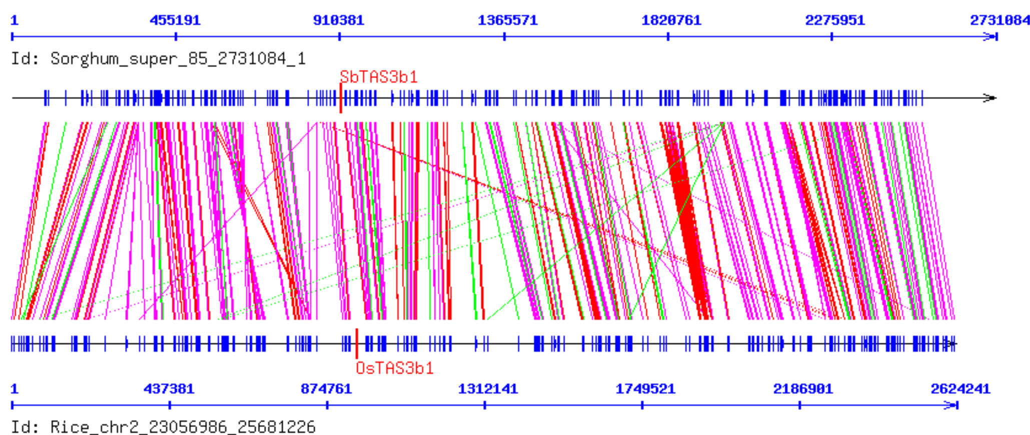


引自 Guo 等（2008b）

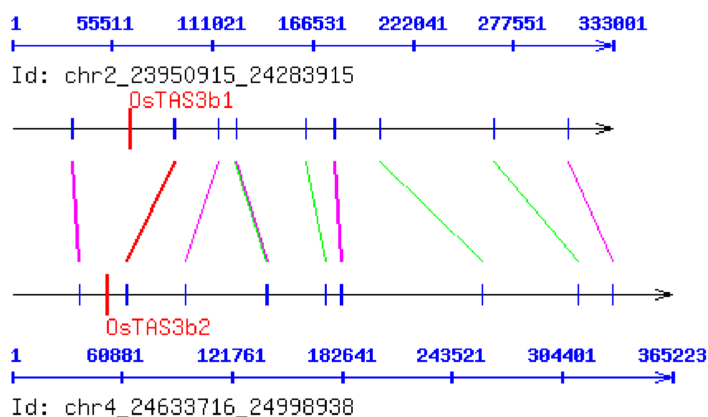
图 7 水稻 miR397 靶基因进化（A）及其结合位点的序列突变情况（B）

ta-siRNA (trans acting siRNAs)是植物中发现的一类 siRNA 基因(*TAS*)，其在 miR390 等的辅助下，调控生长素相关基因 ARF(auxin response factor)，在植物生长发育过程中发挥重要调控功能。目前已在拟南芥中发现四个亚家族（*TAS1-4*）,其中 TAS3 在植物界是保守的。通过保守序列片段，克隆测序和生物信息学方法发现了 51 个来自禾本科的 *TAS3* 基因（Shen et al，2009）。通过序列比较等，发现 *TAS3* 基因通过基因组和单基因倍增，在禾本科基因组中至少有 2 个拷贝，多的可达到近 10 个。水稻基因组倍增而来的 AS3 基因在基因组保持了其共线性关系;同时 TAS3 在不同禾本科基因组上也存在明显的基因组共线性（图 8）。

A



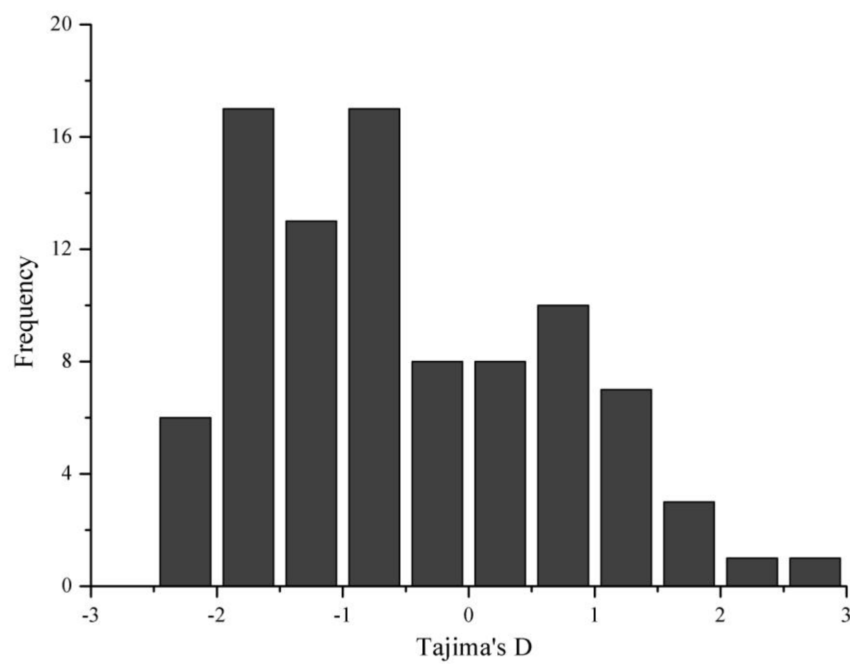
B



引自 Shen 等（2009）

图 8 水稻 ta-siRNA3 （TAS3）基因倍增及其与高粱同源基因的比较基因组学分析

为了调查模式作物—水稻中 miRNA 等小 RNA 是否经受人工选择即驯化的影响，我们对水稻 miRNA 等进行了大规模的群体调查。对 40 个 miRNA 家族的 97 个成员位点进行了重测序，调查群体包括 33 个水稻籼粳亚种。结果表明，与拟南芥的群体调查结果一致，在 miRNA 成熟位点其核苷酸多态性明显低于两端序列，暗示了 miRNA 通过序列互补结合靶基因功能限制的存在。同时，对于保守的 miRNA 家族，其整体的 DNA 多态性相较水稻特异的 miRNA 来说要低一倍，由于保守 miRNA 一般参与基础的代谢网络的调控，因而有可能遭受更强的净化选择而保持序列的保守性(Wang et al. 2010)。另外，我们还对 Tajima *D* 检验显著的 miRNA 位点进行了进一步的正向选择信号的调查。对相应的 miRNA 位点在更大栽培群体（55 个品种）和普通野生稻群体（*O. rufipogon*； 15 个材料）进行重测序用于中性检验等分析，结合 *D* 检验和 HKA 检验的结果，我们找到了几个 miRNA 位点在驯化过程中可能经历了正向选择作用。以 miR390 为例，其调控基因为另一类小 RNA（*TAS3* 基因），中性检验的信号表明 miR390 可能由于选择作用的影响而维持了其特异的调控作用，是水稻驯化和育种选择的直接靶基因（对象）。



引自 Wang 等（2010）

图 9 水稻小 RNA 基因进化选择检测结果。图中包括 94 个位点中性测验 *D* 测验结果的分布。

主要参考文献：

Ashikawa I. Gene-associated CpG islands in plants as revealed by analyses of genomic sequences. **Plant J.**, 2001, 26: 617-625.

Bennetzen J F. Mechanisms and rates of genome expansion and contraction in flowering plants. **Genetica**, 2002, 115: 29-36.

Bennetzen J L, Freeling M. Grasses as a single genetic system-genome composition, colinerarity and compatibility. **Trends Genet.**, 1993, 9: 259–261.

Bennetzen J L. Comparative sequence analysis of plant nuclear genomes: microcolinearity and its many exceptions. **Plant Cell**, 2000, 12: 1021-1029.

Bernardi G. Isochores and the evolutionary genomics of vertebrates. **Gene**, 2000, 241: 3-17.

Betran E, Long M. Expansion of genome coding region by acquisiton of new genes. **Genetica**, 2002, 115: 65-80.

Blanc G, Wolfe K H. Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. **Plant Cell**, 2004, 16: 1667-1678.

Bowers J E, Chapman B A, Rong J, et al. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. **Nature**, 2003, 422: 433–438.

Carels N, Bernardi G. Two classes of genes in plants. **Genetics**, 2000, 154: 1819-1825.

Devos K M, Brown J K M, Bennetzen J F. Genome size reduction through illegitimate recombination counteracts genome expansion in *Arabidopsis*. **Genome Res.**, 2002, 12: 1075-1079.

Fan L, Xu G, Zhang Y and Guo X. Duplication of rice (*Oryza sativa*) genome [A]. Proceedings of 8th National Systematic and Evolutionary Botany Youth Seminar and Systematic & Evolutionary Biology Conference, China, 7.20~7.25. 2004.

Feng Q, Zhang Y, Hao P, et al. Sequence and analysis of rice chromosome 4. **Nature**, 2002, 420: 316–320.

Freeling M. Grasses as a single genetic system. **Plant Physiol.**, 2001, 125: 1191–1197.

Fujimori S, Washio T, Higo K, et al. A novel feature of microsatellites in plants: a distribution gradient along the direction of transcription. **FEBS Lett.**, 2003, 554: 17-22.

Gaut B S, Doebley J F. DNA sequence evidence for the segmental allotetraploid origin of maize. **Proc. Natl. Acad. Sci. USA**, 1997, 94: 6809-6814.

Gaut B S. Evolutionary dynamics of grass genomes. **New Phytologist**, 2002, 154:15-28.

Goff S A, Rick D, Lan T H, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). **Science**, 2002, 296, 92-100.

Grover C E, Kim H, Paterson A H, et al. Incongruent patterns of local and global genome size evolution in cotton. **Genome Res.**, 2004, 14: 1474-1482.

Guo Xingyi, Guohua Xu, Yang Zhang, Xiao Wen, Weimin Hu, Longjiang Fan. Incongruent evolution of chromosome size in rice. **Genetics and Molecular Research** 2006, 5(2): 373-389

Guo Xingyi, Jiandong Bao, Longjiang Fan. Evidence of selectively driven codon usage in rice: Implications for GC content evolution of Gramineae Genes. **FEBS Letters** 2007, 581(5): 1015-1021

Guo Xingyi, Songlin Ruan, Weiming Hu, Daguang Cai, Longjiang Fan. Chloroplast DNA insertions into the nuclear genome of rice: the genes, sites and ages of insertion involved. **Funct Integr Genomics**, 2008a, 8:101–108.

Guo Xinyi, Yijie Gui, Yu Wang, Qian-Hao Zhu, Chris Helliwell and Longjiang Fan. Selection and mutation on microRNA target sequences during rice evolution. **BMC Genomics**, 2008b, 9:454

Guyot R, Keller B. Ancestral genome duplication in rice. **Genome**, 2004, 47(3): 610-4.

Han B, Xue Y. Genome-wide intraspecific DNA-sequence variations in rice. **Curr. Opin. Genet. Dev.**, 2003, 13: 134-138.

Hiratsuka J, Shimada H, Whittier R, et al. The complete sequence of the rice (*Oryza*

sativa) chloroplast genome: intermolecular recombination between distinct tRNA genes accounts for a major plastid DNA inversion during the evolution of the cereals. **Mol. Gen. Genet.**, 1989, 217: 185–194.

Hurst L D. The Ka/Ks ratio: diagnosing the form of sequence evolution. **Trends Genet.**, 2002, 18: 486-487.

Ilic K, SanMiguel P J, Bennetzen J L. A complex history of rearrangement in an orthologous region of the maize, sorghum, and rice genomes. **Proc. Natl. Acad. Sci. USA**, 2003, 100: 12265–12270.

Ma J, Bennetzen J L. Rapid recent growth and divergence of rice nuclear genomes [J]. **Proc. Natl. Acad. Sci. USA**, 2004, 101: 12404-12410.

Ma J, Devos K, Bennetzen J L. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. **Genome Res.**, 2004, 14: 860-869.

Notsu Y, Masood S, Nishikawa T, et al. The complete sequence of the rice (*Oryza sativa* L.) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution of flowering plants. **Mol. Genet. Genomics**, 2002, 268: 434–445.

Paterson A H, Bowers J E, Chapman B A. Ancient polyploidization predating divergence of the cereals, and its consequences for comparative genomics. **Proc. Natl. Acad. Sci. USA**, 2004, 101: 9903-9908.

Paterson A H, Bowers J E, Peterson J, et al. Structure and evolution of cereal genomes. **Curr. Opin. Genet. Dev.**, 2003, 13: 644–650.

Petrov D A. DNA loss and evolution of genome size in *Drosophila*. **Genetica**, 2002, 115: 81-91.

Sall T, Jakobsson M, Lind-hallden C, et al. Chloroplast DNA indicates a single origin of the allotetraploid *Arabidopsis suecica*. **J. Evol. Biol.**, 2003, 16: 1019-1029.

SanMiguel P, Tikhonov A, Jin Y-K, et al. Nested retrotransposons in the intergenic regions of the maize genome. **Science**, 1996, 274: 765–768.

Sasaki T, Matsumoto T, Yamamoto K, et al. The genome sequence and structure of rice chromosome 1. **Nature**, 2002, 420: 312–316.

Shahmuradov I A, Akbarova Y Y, Solovyev V V, et al. Abundance of plastid DNA

insertions in nuclear genomes of rice and *Arabidopsis*. **Plant Mol. Biol.**, 2003, 52: 923–934.

Shen Dan, Wang Sheng, Chen Huan, Fan Longjiang. Molecular phylogeny of miR390-guided *trans*-acting siRNA genes (*TAS3*) in the grass family. **Plant Systematics Evolution**, 2009, in press

Shirasu K, Schulman A H, Lahaye T, et al. A contiguous 66-kb barley DNA sequence provides evidence for reversible genome expansion. **Genome Res.**, 2000, 10: 908–915.

Simillion C, Vandepoele K, Van Montagu M C E, et al. The hidden duplication past of *Arabidopsis thaliana*. **Proc. Natl. Acad. Sci. USA**, 2002, 99: 13627–13632.

Song R, Llaca V, Messing J. Mosaic organization of orthologous sequences in grass genomes. **Genome Res.**, 2003, 12: 1549-1555.

Swanson W J. Adaptive evolution of genes and gene families. **Curr. Opin. Genet. Dev.**, 2003, 13: 617–622.

The Rice Chromosome 10 Sequencing Consortium. In-depth view of structure, activity, and evolution of rice chromosome 10. **Science**, 2003, 300: 1566-169.

Vandepoele K, Simillion C, Van de Peer Y. Evidence that rice and other cereals are ancient aneuploids **Plant Cell**, 2003, 15: 2192–2202.

Vicient C M, Suoniemi A, Anamthawas-Jonsson K, et al. Retrotransposon BARE-1 and its role in genome evolution in the genus Hordeum. **Plant Cell**, 1999, 11: 1769–1784.

Wang X, Shi X, Hao B, et al. Duplication and DNA segmental loss in the rice genome: implications for diploidization. **New Phytologist**, 2005, 165: 937-946.

Wang Sheng, Qianhao Zhu, Xingyi Guo, Yijie Gui, Jiandong Bao, Chris Helliwell, Longjiang Fan. Molecular evolution and selection of a gene encoding two tandem microRNAs in rice. **FEBS Letters**, 2007, 581:4789-4793.

Wang Y, Shen D, Bo S, Chen H, Zheng J, Zhu QH, Helliwell C, Fan L. Sequence variation and selection of small RNAs in domesticated rice. **BMC Evol Biol**, 2010, Revised

Wendel J F. Genome evolution in polyploids. **Plant Mol. Biol.**, 2000, 42: 225-249.

Wolfe K H, Gouy M, Yang Y-W, et al. Date of the monocot-dicot divergence estimated from chloroplast DNA sequence data. **Proc. Natl. Acad. Sci. USA**, 1989, 86: 6201-6205.

Wong G K, Wang J, Tao L, et al. Compositional Gradients in Gramineae Genes. **Genome Res.**, 2002 12: 851-856.

Yu J, Hu S, Wang J, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). **Science**, 2002, 296: 79-92.

Yu J, Wang J, Lin W, et al. The genome of *Oryza sativa*: A history of duplication. **PloS. Biol.**, 2005, 3(2): e38.

Zhang L, Pond K S, Gaut B S. A survey of the molecular evolutionary dynamics of twenty-five multigene families from four grass taxa. **J. Mol. Evol.**, 2001, 52: 144–156.

Zhang Y, Xu G H, Guo X Y, Fan L. Two ancient rounds of polyploidy in rice genome. **Journal of Zhejiang University SCIENCE**, 2005, 6(2): 87-90.

Zhu Qian-Hao, Andrew Spriggs, Louisa Matthew, Longjiang Fan, Gavin Kennedy, Frank Gubler, and Chris Helliwell. A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. **Genome Res**, 2008, 18:1456-1465.

第五章 分子进化：系统树的构建[1]

自 20 世纪中叶，随着分子生物学的不断发展，进化研究也进入了分子进化(molecular evolution)研究水平，并建立了一套依赖于核酸、蛋白质序列信息的理论和方法。随着基因组测序计划的实施，基因组的巨量信息对若干生物领域重大问题的研究提供了有力的帮助，分子进化研究再次成为生命科学中最引人注目的领域之一。这些重大问题包括：遗传密码的起源、基因组结构的形成与演化、进化的动力、生物进化等等。分子进化研究目前更多地是集中在分子序列上，但随着越来越多生物基因组的测序完成，从基因组水平上探索进化奥秘，将开创进化研究的新天地。人与老鼠的基因组大小相似，都含有约 30 亿碱基对，基因的数量也相近，可人与老鼠为何差异如此之大？从进化的角度如此解释？是否可以在浩如烟海的基因组密码中获得答案？

第一节 系统树及其它

一．系统树

分类学涉及的问题是将生物合理地分成一定的类群，使类群内的个体成员相同或非常相似。分类学可以进行物种的分类。对于进化研究，分类涉及到系统发育的重构(reconstruction of phylogenies)，构建系统发育过程有助于通过物种间隐含的种系关系揭示进化动力的实质。Nei(1987)、Li 和 Graur(1991)等人已对构建系统发育过程进行了全面的总结，本章只提示性地介绍相关方法。

表型的(phenetic)和遗传的(cladistic)数据有着明显差异。Sneath 和 Sokal(1973)将表型性关系定义为根据物体一组表型性状所获得的相似性，而遗传性关系含有祖先的信息，因而可用于研究进化的途径。这两种关系可用于系统树(phylogenetic tree)或树状图(dendrogram)来表示。表型分枝图(phenogram)和进化分枝图(cladogram)两个术语已用于表示分别根据表型性的和遗传性的关系所建立的关系树。进化分枝图可以显示事件或类群间的进化时间，而表型分枝图则不需要时间概念。在本章，我们将不会十分注重这一区别，正如 Nei(1987)指出的，如果表型相似性的尺度意味着进化上的相似性的程度，则有关表型的方法就可以提供遗传上的关系树。文献中，更多地是使用“系统树”一词来表示进化的途径，另外还有系统发育树、物种树(species tree)、基因树等等一些相同或含义略有差异的名称。

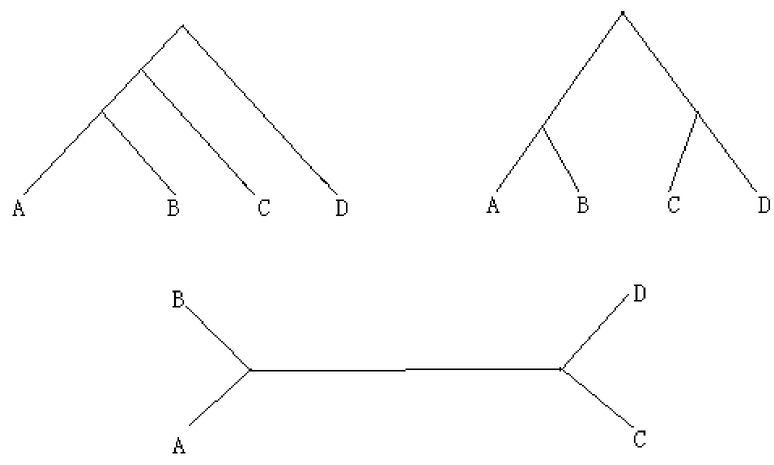
系统树分有根(rooted)和无根(unrooted)树。图 5.1 中显示了 4 个物种部分有根树和无根树形式。有根树反映了树上物种或基因的时间顺序，而无根树只反映分类单元之间的距离而不涉及谁是谁的祖先问题。

用于构建系统树的数据有二种类型：一种是特征数据(character data)，它提供了基因、个体、群体或物种的信息；二是距离数据(distance data)或相似性数据(similarity data)，它涉及的则是成对基因、个体、群体或物种的信息。距离数据可由特征数据计算获得，但反过来则不行。这些数据可以矩阵的形式表

---

[1]本部分内容主要取自Weir B. S.（徐云碧等译）. 遗传学数据分析—群体遗传学离散型数据分析方法，北京：中国农业出版社，1996

达。距离矩阵(distance matrix)是在计算得到的距离数据基础上获得的，距离



的计算总体上是要依据一定的遗传模型，并能够表示出两个分类单位间的变化量。系统树的构建质量依赖于距离估算的准确性。

图 5.1 4 个物种(A、B、C 和 D)的 2 种有根树和 1 种无根树形式

系统树的构建主要有三种方法。距离矩阵法(distance matrix method)是根据每对物种之间的距离，其计算一般很直接，所生成的树的质量取决于距离尺度的质量。距离通常取决于遗传模型。最大简约(maximum parsimony)法较少涉及遗传假设，它通过寻求物种间最小的变更数来完成的。对于模型的巨大依赖性是最大似然(maximum likelihood)法的特征，该方法在计算上繁杂，但为统计推断提供了基础。

二．遗传模型和序列距离

遗传模型在系统树构建中非常重要，因为距离计算过程必须在一定的遗传假设下才可能进行。以下以两个在 DNA 序列距离计算中最为常用的遗传模型为例，说明距离数据的计算由来。

在分子进化研究中，我们往往认定这样的一个假设，即序列是同源的，它们具有单一祖先序列；这一祖先序列在进化过程中发生了一系列的核苷酸突变。图 5.2 表示了各种核苷酸变化情况。

在以上的假设基础上，Judes 和 Cautor 进一步假设每一碱基具有同等机率突变为另外 3 种碱基中的任何一种，其频率常数为μ/3，μ为碱基替换频率。Kimura(1980)考虑到转换(transition，两种嘧啶或两种嘌呤碱基之间的突变)和颠换(transversion，一个嘧啶和一个嘌呤碱基之间的突变)具有不同的频率，a 和β。表 5.1 简要说明了以上两种遗传模型。

表 5.1 Judes-Cantor 单参数模型(上三角部分)和 Kimura 两参数模型(下三角部分)。a、β分别为两种碱基间 2 个不同的置换频率。

	A	T	G	C

A		α	α	α
T	β		α	α
G	α	β		α
C	β	α	β	

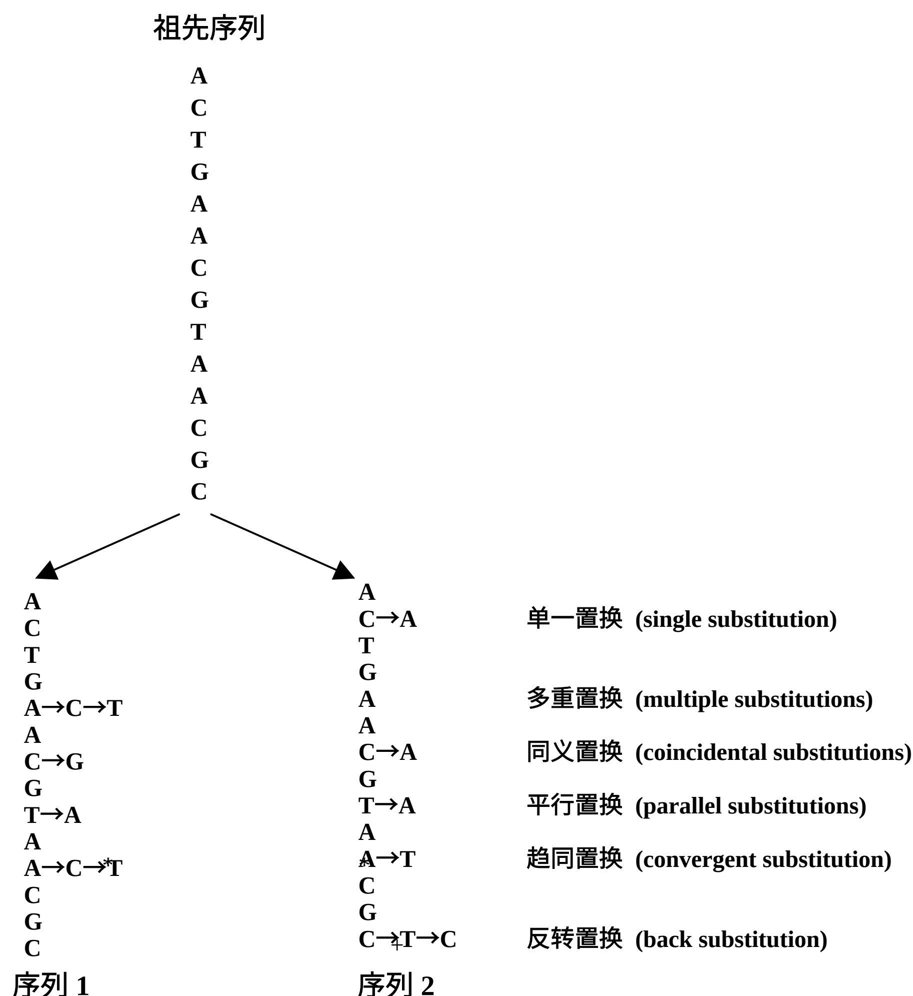


图 5.2 同源序列间的核苷酸置换(Li & Graur,1991)

根据以上遗传模型，Judes 和 Cantor(1969)提出了 DNA 序列距离 *K*(最早为氨基酸序列引入)计算公式：

$$K = \frac{3}{4}\ln(\frac{4}{4q-1}) \approx 2\mu t \qquad (5.1)$$

其中 q 为同源 DNA 序列中具有相同碱基的概率，经过 t 世代，由于祖先序列的趋异变化，其值为：

$$q_t = \frac{1}{4} + \frac{3}{4}(1 - \frac{8\mu}{3})^t \qquad (5.2)$$

μ为碱基替换频率。

距离 K 适用于显示两条序列从一个祖先序列趋异进化以来的时间,并能用于序列间系统树的构建。在计算时，均需要将序列作初步的列线分析。Kimura 在其两参数模型下证实，由于趋异变化，由转换造成差异(Ⅰ型变化)或由颠换造成差异(Ⅱ型变化)的碱基，随时间而变化：

$$P_{It} = \frac{1}{4}(1 - 2e^{-4(\alpha+\beta)t} + e^{-8\beta t})$$

$$P_{It} = \frac{1}{4}(1 - e^{-8\beta t}) \qquad (5.3)$$

如果 k＝ａ＋2β是单位时间碱基替换的总频率，则适合作为系统树的距离尺度为：

$$K = -\frac{1}{2}\ln[(1-2p_I - p_{II})\sqrt{1-2P_{II}}] \approx 2kt \qquad (5.4)$$

该类距离可用于有关系统树距离矩阵中，用样本比值代入（5.4）式就可估计这些距离。

Kimura 以兔和鸡的β-球蛋白序列为例(见图 5.3)，计算了上述距离。序列长 438bp，有 58 个Ⅰ型变化、63 个Ⅱ型变化。因此，$\tilde{p}_I$=0.1324，$\tilde{P}_{\Pi}$=0.1438，Kimura 距离为 0.3513。这与只根据相同碱基比例$\tilde{q}$=0.7237 所得 Jukes-Cantor 距离 0.3446 没有本质上的差异。

图 5.3 兔和鸡的β-球蛋白序列。每两条序列上下两行星号表示由转换（Ⅰ型变化)或颠换(Ⅱ型变化)造成的碱基差异。

DNA序列距离K又可称为DNA序列间的分歧度(sequence divergence)，即序列间相异性的一个指标。蛋白质序列的分歧度分为两序列同义变化的分歧度(K_S)和非同义变化的分歧度(K_A)，根据Jukes-Cautor单参数模型和Kimura两参数模型等遗传模型，可以分别计算得到两序列的分歧度(或称为蛋白质序列间的距离)。

三．分子进化与系统发育分析软件

软件名称	网址	说　明
PHYLIP	http://evolution.genetics.washinton.edu/phylip/software.html	目前发布最广，用户最多的通用系统树构建软件，由美国华盛顿大学 Felsenstein 开发，可免费下载，适用绝大多数操作系统
PAUP	scavotto@sinauer.com 或 ftp://onyx.si.edu/paup	国际上最通用的系统树构建软件之一，美国 simthsonion institute 开发，仅适用 Apple-Macintosh 和 UNIX 操作系统
Tree of Life	http://phylogeny.arizona.edu/tree/program/program.html	美国 University of Arizona 建立的系统发育方面网站
MEGA	http://bioinfo.weizmann.ac.il/databases/info/mega.sof	美国宾西法尼亚州立大学 MasatoshiNei 开发的分子进化遗传学软件
MOLPHY	ftp://ftp sunmh.ism.ac.jp/pub/molphy	日本国立统计数理研究所开发，最大似然法构树
PAML	http://abacus.gene.ucl.ac.uk/software/paml.html	英国 University college London 开发，最大似然法构树和分子进化模型
PUZZLE	ftp://fx.zi.biologie.uni-muenchen.de/pub/puzzle	应用 quarter puzzling 方法(一种最大简约法)构建系统树
TreeView	http://taxonomy.zoology.gla.ac.uk/rod/treeview.html	英国 University of Glasgow 开发
phylogeny	http://www.ebi.ac.uk/biocat/phylogeny.html	欧洲生物信息研究所(EBI)的系统发育分析软件

第二节　　距离矩阵法

系统树可建立在(遗传)距离矩阵的基础上。这里的遗传距离为所有成对实用分类单位(operational taxonomic units，OTU)之间的距离。对于 t 个 OUT，每一对之间的距离矩阵列于表 5.2。

表 5.2　 t 个实用分类单位（OUT）间的距离矩阵

		OUT 数				
		1	2	3	…	t
OUT 数	1	-	d_{12}	d_{13}	…	d_{1t}
	2	d_{21}	-	d_{23}	…	d_{2t}
	3	d_{31}	d_{32}	-	…	d_{3t}
	…	…	…	…	…	…
	t	d_{t1}	d_{t2}	d_{t3}	…	-

用这些距离对 OUT 进行表型意义的分类可借助于聚类分析(clustering)，聚类过程可以看作是鉴别具有相近 OUT 类群的过程。

一．平均连接聚类法(UPGMA 法)

可以采用几种聚类方法，这些方法包括序贯法(sequential)．聚合法(agglomerative)．分层法(hierarchical)和非重叠法(nonoverlapping)等。应用最广泛的是平均连接聚类法(average linkage clustering)或称为 UPGMA 法

(应用算术平均数的非加权成组配对法，unweighted pair-group method using an arithmetic average)。该法将类间距离定义为两个类的成员所有成对距离的平均值。

作为实例，我们考虑图 5.4 所列的线粒体 DNA 序列的资料。每对序列间的 Jukes-Cantor 距离取决于每对序列间差异核苷酸的观察数。如果在两条序列中相同碱基的比例为 q，则距离 K 可估计为

$$\tilde{K} = \frac{3}{4}\ln(\frac{3}{4q-1})$$

序列的差异和距离列于表 5.3

1．人类　GTAAATATAG TTTAACCAAA ACATCAGATT GTGAATCTGA CAACAGAGGC TTACGACCCC TTATTTACC
2．黑猩猩　GTAAATATAG TTTAACCAAA ACATCAGATT GTGAATCTGA CAACAGAGGC TCACGACCCC TTATTTACC
3．大猩猩　GTAAATATAG TTTAACCAAA ACATCAGATT GTGAATCTGA TAACAGAGGC TCACAACCCC TTATTTACC
4．猩猩　GTAAATATAG TTTAACCAAA ACATTAGATT GTGAATCTAA TAATAGGGCC CCACAACCCC TTATTTACC
5 长臂猿　GTAAACATAG TTTAATCAAA ACATTAGATT GTGAATCTAA CAATAGAGGC TCGAAACCTC TTGCTTACC

图 5.4 五种生物线粒体 DNA 序列

最近的距离是人类和黑猩猩之间的，将它们合并为一个类。其它序列与这个新类之间的距离就是该序列到新类各成员间的平均距离：

$$d_{(hu-ch),go} = \frac{1}{2}(d_{hu,go} + d_{ch,go}) = 0.037$$

$$d_{(hu-ch),or} = \frac{1}{2}(d_{hu,or} + d_{ch,or}) = 0.135$$

$$d_{(hu-ch),gi} = \frac{1}{2}(d_{hu,gi} + d_{ch,gi}) = 0.189$$

表 5.3　图 5.4 中 5 个线粒体序列的差异核苷酸数(对角线下)和 Jukes-Cantor 距离(对角线上)

	人类(hu)	黑猩猩(ch)	大猩猩(go)	猩猩(or)	长臂猿(gi)
人类(hu)	-	0.015	0.045	0.143	0.198
黑猩猩(ch)	1	-	0.030	0.126	0.179
大猩猩(go)	3	2	-	0.092	0.179
猩猩(or)	9	8	6	-	0.179
长臂猿(gi)	12	11	11	11	-

图 5.4

距离矩阵可简缩为：

	(hu-ch)	go	or	gi
hu-ch		0.037	0.135	0.189
go				0.179
or				0.179
gi				

其中人类－黑猩猩（hu-ch）与大猩猩（go）之间的距离最小。将它们合并为一类。新距离为：

$$d_{(hu-ch-go),or} = \frac{1}{3}(d_{hu,or} + d_{ch,or} + d_{gp.pr}) = 0.121$$

$$d_{(hu-ch-go),gi} = \frac{1}{3}(d_{hu,gi} + d_{ch,gi} + d_{go,gi}) = 0.185$$

下一个简缩后的距离矩阵为：

	(hu-ch-go)	or	gi
(hu-ch-go)		0.121	0.185
or			0.179
gi			

现在人类－黑猩猩－大猩猩（hu-ch-go）和猩猩（or）之间的距离最小，将其并为一类，从该四合体到猩猩序列的距离为：

$$d_{(du-ch-go-or),gi} = \frac{1}{4}(d_{hu,gi+dch,gi} + d_{go,gi} + d_{or,gi}) = 0.183$$

上述聚类结果可表示为图 5.5 所示的树状图。在构建树状图时，分枝点安置在两个序列或类的中点。图中成对序列间的距离为分枝长度之和。

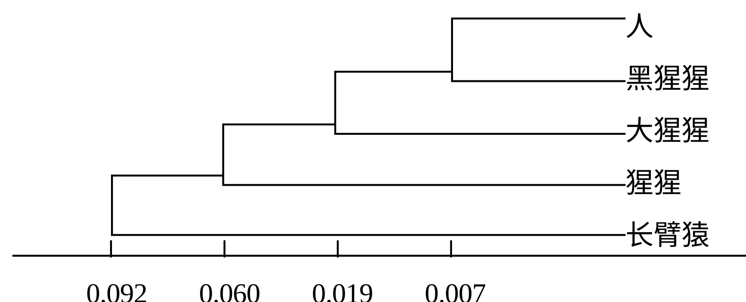


图 5.5 平均连接聚类法系统树

UPGMA 方法广泛用于距离矩阵。Nei 等（1983）模拟了构建树的不同方法，发现当沿树上所有分枝的突变率相同时， UPGMA 法一般能够得到较好的结果。但必须强调有关突变率相等(或几乎相等)的假设对于 UPGMA 的应用是重要的。另一些模型研究(如 Kim 和 Burgman，1988)已证实当各分枝的突变率不相等时，这一方法的结果不尽人意。当各分枝突变率相等时，认为分子钟(molecular clock)在起作用。

二．Fitch-Margoliash 算法

UPGMA 法包含这样的假定：沿着树的所有分枝突变率为常数。Fitch 和 Margoliash(1967)所发展的方法去除了这一假定。该法的应用过程包括插入“丧失的”OUT 作为后面 OUT 的共同祖先，并每次使分枝长度拟合于 3 个 OTU 组。现在用图 5.4 的线粒体资料来说明 Fitch-Margoliash 法则。

将 OUT 分为三组：距离最近的一对为 A=人类(hu)和 B=黑猩猩(ch)，剩下 X=(大猩猩 go，猩猩 or，长臂猿 gi)。引入树节 C 作为 A 和 B 的直接祖先。设从 C 到 A、B 的长度为 a、b，从 C 到 X 的为 x(图 8.4)。A、B、C 之间的 3 个成对距离提供了可解 3 个未知数的 3 个方程：

$$\begin{cases} a + x = d_{AX} = d_{AB} = \frac{1}{3}(0.045 + 0.143 + 0.198) = 0.129 \\ b + x = d_{BX} = d_{B\bar{A}} = \frac{1}{3}(0.030 + 0.126 + 0.179) = 0.112 \\ a + b = d_{AB} = 0.015 \end{cases}$$

设定如下符号约定：设d_{UV}为节点U到节点V的距离，$d_{U\bar{V}}$为节点U到V外所有节点的平均距离，d_{U^*V}为U以下所有末端节到V的平均距离。U*表示从同一字母的节点U下的一组末端树节。

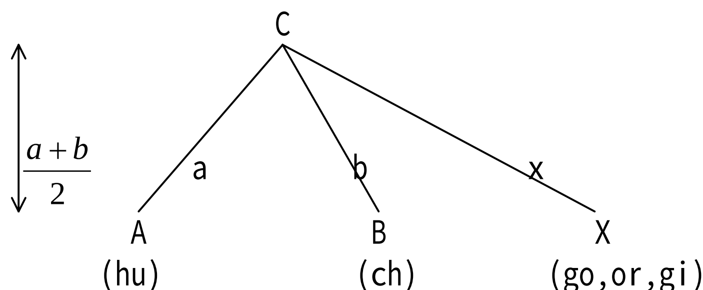


图 5.6 将 Fitch-Margoliash 算法应用于图 5.4 线粒体资料的初始步骤

第一个方程采用了从 A 到 X 的每一成员的平均距离。解以上三个方程得：

a=0.016，b=-0.001

为了方便以见，负的值定为 0，因此 b=0。a、b 的平均值为树节 C 的高度，该值为 0.008。

用C代替A、B，按UPGMA所采用的方式再计算距离值，得到下一个最近的一对为C和D(=go)。引入树节E作为C和D的直接祖先。如图 5.7 所示，节点C*和E、D和E，E和X的分枝长度分别为c、d和x。现在X只包含猩猩(or)和长臂猿(gi)。要解的 3 个方程为：

$$\begin{cases} c+d=d_{C^*D}=\frac{1}{2}(0.045+0.030)=0.037 \\ c+x=d_{c^*x}=d_{(AB)^*}=\frac{1}{4}(0.143+0.198+0.126+0.179)=0.162 \\ d+x=d_{DX}=\frac{1}{2}(0.092+0.179)=0.136 \end{cases}$$

因此

c=0.032， b=0.006

节点E的高度为(c+d)/2=0.019。由于c度量了C到E距离以及从A和B到C的平均距离，所以c减去树节C的高度就得到C到E之间的分枝长度c'。换言之

c'=0.032-0.008=0.024

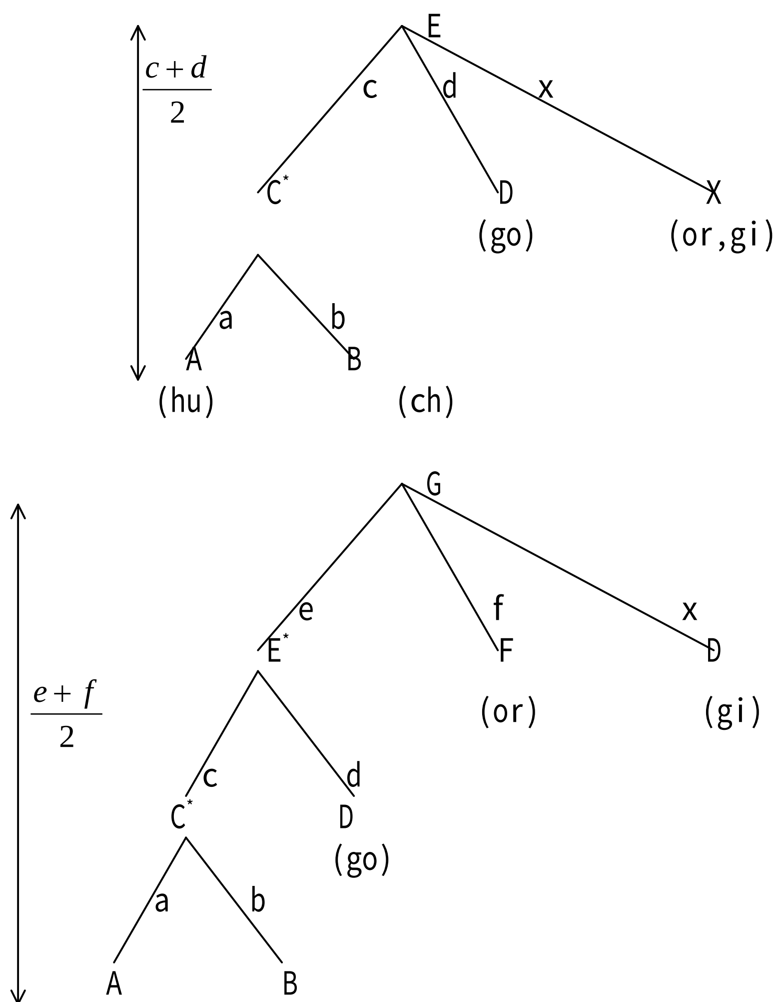


图 5.7 将 Fitch-Margoliash 算法应用于图 5.4 线粒体序列资料时的中间步骤

随着 OUT 简缩到 E、猩猩(or)和长臂猿(gi)。距离最近的一对就是 E 和 F(=or)了。引入 G 作为直接祖先，余下的 X=gi。要得到分枝长度所要解的方程为

$$\begin{cases} e + f = d_{E^*F} = \frac{1}{3}(0.143 + 0.126 + 0.092) = 0.121 \\ e + x = d_{E^*X} = \frac{1}{3}(0.198 + 0.179 + 0.179) = 0.185 \\ f + x = d_{FX} = 0.179 \end{cases}$$

故

e=0.063, f=0.057

节点G的高度为(e+f)/2=0.060，从E到G的分枝长度e′为e与E的高度之差，即0.063-0.019=0.044。

Fitch-Margoliash 算法计算过程可以到此为止,图 5.8 给出了其无根系统树。

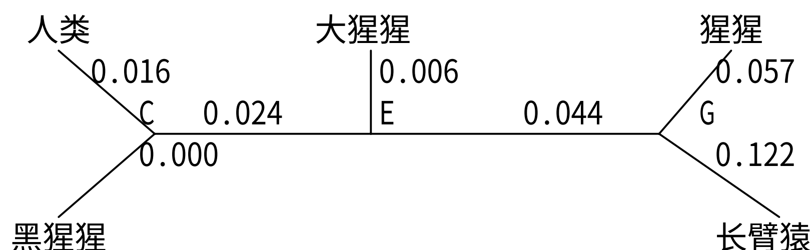


图 5.8 图 5.4 所列线粒体序列资料的 Fitch-Margoliash 无根系统树

如果不假定沿所有分枝具有相同的变更率,则由 Fitch-Margoliash 算法只能得到无根系统树。如果设置树根 I，并假定从 I 到现在所有序列的两个分枝具有相等的变更率，因而从 G 到 I 的距离 g 与从 H 到 I 的距离 h 是相等的，则有根树就可以采用与 UPGMA 提供的相同拓扑方法来获得。由于

$$g + h = d_{G^*H}$$

$$= \frac{1}{4}(0.198 + 0.179 + 0.179 + 0.179) = 0.184$$

所以g=h=0.092,且从G到I的距离g' 为g减去G的高度，即 0.032。将所有这些分枝长度一起考虑便得到图 5.9 所示有根系统树。

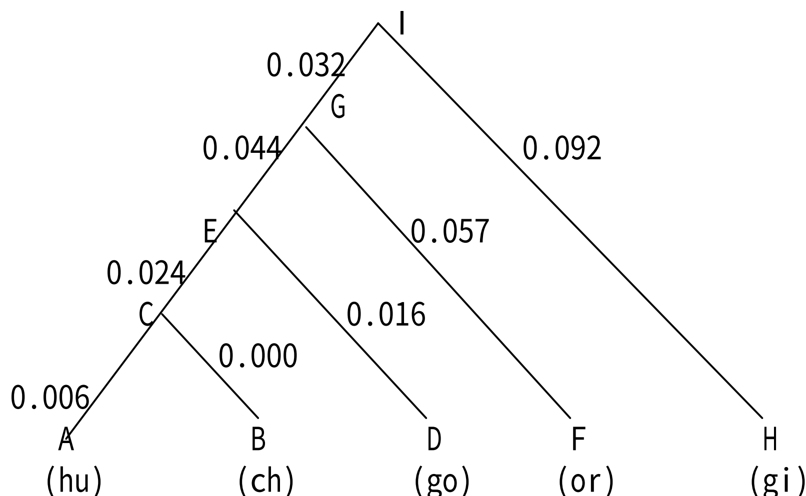


图 5.9 图 5.4 所列线粒体序列资料的 Fitch-Margoliash 有根树状图

Fitch和Margoliash承认他们的法则所得到的拓扑结构可能是不正确的，并建议考查其它的拓扑结构。可以采用Fitch和Margoliash(1967)称之为“ 百分标准差 ”的一种拟合优度来比较不同的系统树，最佳系统树应具有最小的百分标准差。如果d_{ij}为n个OUT中i和j的观测距离(即Judes-Cantor距离)，e_{ij}为i和j之间分枝长度之和，则

$$s = \left\{ \frac{\sum [(d_{ij} - e_{ij}) / d_{ij}]^2}{n(n-1)} \right\}^{\frac{1}{2}} \times 100 \qquad (5.5)$$

为百分标准差。考虑到可加性的假定，因而有任意两个节点之间的距离就是它们之间分枝长度之和。对于图 5.7 的系统树，观测距离和分枝长度列于表 5.4，其百分标准差为 1.94。通过调整适合系统树的分枝长度来降低 s 是可能的。

根据百分标准差选择系统树，其最佳系统树可能与由 Fitch-Margoliash 法则所得的不相同。当存在分子钟时，可以预期这一标准差的应用将给出类似于 UPGMA 方法的结果。如果不存在分子钟，因而在不同的世系(分枝)中的变更率是不同的，则 Fitch-Margoliash 标准就会比 UPGMA 好得多。

表 5.4 图 5.4 中 5 种线粒体序列的观测距离(对角线上)和采用 Fitch-Margoliash 算法计算所得距离(对角线下)

	人类	黑猩猩	大猩猩	猩猩	长臂猿
人类	—	0.015	0.045	0.143	0.198
黑猩猩	0.016	—	0.030	0.126	0.179
大猩猩	0.046	0.030	—	0.092	0.179
猩猩	0.141	0.125	0.107	—	0.179
长臂猿	0.208	0.192	0.174	0.181	—

通过选择不同的 OUT 作为初始配对单位，就可以选择其它的系统树进行考查。具有最低百分标准差的系统树即被认为是最佳的，并且这个标准是建立在应用 Fitch-Margoliash 算法的基础上的。例如，首先将人类和大猩猩分为一类，然后依次将黑猩猩、猩猩和长臂猿增加进去。但是，在这种情况下，第二个内部节点 E 的高度低于第一个内部节点 C 的高度，观测距离和计算距离之间的适合度就不如第一种情形那么好。

三．邻接法

邻接法(Neighbor-joining Method)由 Saitou 和 Nei(1987)提出。该方法通过确定距离最近(或相邻)的成对分类单位来使系统树的总距离达到最小。相邻是指两个分类单位在某一无根分叉树中仅通过一个节点(node)相连。图 5.2 中，人与黑猩猩是相邻的，人与大猩猩则不是；如果人与黑猩猩组成一个新类，则该新类与大猩猩又成为相邻。总之，通过循序地将相邻点合并成新的点，就可以建立一个相应的拓扑树。

邻接法的一般步骤：

①计算第i终端节点(即分类单位i)的净分歧度r_i

$$r_i = \sum_{k=1}^{N} d_{ik} \tag{5.6}$$

其中N为终端节点数，d_{ik}为节点i和节点k之间的距离，有$d_{ik}=d_{ki}$

②计算并确定最小速率校正距离(rate-corrected distance)M_{ij}：

$$M_{ij} = d_{ij} - \frac{r_i + r_j}{N-2} \tag{5.7}$$

③定义一个新节点 u，u 节点由节点 i 和 j 组合而成。节点 u 与节点 i 和 j 的距离为：

$$S_{iu}=\frac{d_{ij}}{2}+\frac{r_i+r_j}{2(N-2)}$$

$$S_{ju}=d_{ij}-S_{iu} \tag{5.8}$$

节点 u 与系统树其它节点 k 的距离为：

$$d_{ku}=\frac{d_{ik}+d_{jk}-d_{ij}}{2} \tag{5.9}$$

④从距离矩阵中删除列节点 i 和 j 的距离，N 值(总节点数)减去 1
⑤如果尚余 2 个以上终端节点，返回到步骤①继续计算，直至系统树完全建成。

以上每一步可以产生一个中间节点，并最终画出系统树。图中各分枝的角度是随意的。

现还以表 5.3 线粒体序列为例说明以上计算过程。表 5.5 列出了各步计算的结果，其中最小M_{ij}值用星号注明。第一步，星号(or)和长臂猿(gi)之间的M_{ij}值最小，则它们用节点 1 取代，进入第 2 步，则新节点(节点 1)到这二个节点的距离为：

$$d_{or,节点1}=\frac{1}{2}d_{or,gi}+\frac{r_{or}-r_{gi}}{6}=0.057$$

$$d_{gi,节点1}=d_{or,gi}-d_{or,节点1}=0.122$$

节点 1 到其它各节点的距离见表 5.5 第二步矩阵。在该矩阵中，人(hu)和黑猩猩(ch)的M_{ij}值最小，则它们又形成一个新节点(节点 2)……依次类推，便可最终完成矩阵的计算和邻接法无根系统树。

表 5.5 邻接法计算线粒体序列(图 5.4)的距离d_{ij}(上对角线部分)和M_{ij}(下对角线部分)

		hu j=1	ch j=2	go j=3	or j=4	gi j=5	净分歧度 r_i
hu	i=1	0.000	0.015	0.045	0.143	0.198	0.401
ch	i=2	-0.235	0.000	0.030	0.126	0.179	0.350
go	i=3	-0.204	-0.202	0.000	0.092	0.179	0.346
or	i=4	-0.171	-0.171	-0.203	0.000	0.179	0.540
gi	i=5	-0.181	-0.183	-0.181	-0.246	0.000	0.735

		hu j=1	ch j=2	go j=3	节点 1 j=4	r_i
hu	i=1	0.000	0.015	0.045	0.081	0.141
ch	i=2	-0.110	0.000	0.030	0.063	0.108
go	i=3	-0.086	-0.084	0.000	0.046	0.121
节点 1	i=4	-0.085	-0.086	-0.110	0.000	0.190

		go j=1	节点 1 j=2	节点 2 j=3	r_i
go	i=1	0.000	0.046	0.030	0.076
节点 1	i=2	-0.141	0.000	0.065	0.111
节点 2	i=3	-0.141	-0.141	0.000	0.095

		go j=1	节点 3 j=2
go	i=1	0.000	0.005
节点 3	i=2		0.000

*hu、ch、go、or 和 gi 分别代表人、黑猩猩、大猩猩、猩猩和长臂猿

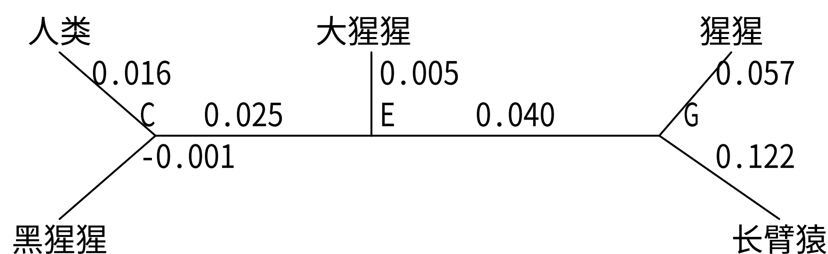


图 5.10 根据线粒体序列(图 5.4)构建邻接法无根系统树

第三节　简约法

简约法(Parsimony)明显注重每一物种观测的特征值，而不是概括特征值之间差异的序列间距离。该法由 Edwards 和 Cavalli-Sforza(1963)以“最小进化原理”的名称应用于基因频率资料。如果有一组物种的序列可供利用，那么连接它们的最为简约的拓扑结构就可能得到。但一般无法获得分枝长度。

对于每种可能的拓扑结构，每一节点的序列就是产生两个直接后裔序列所需变更最小的序列。然后可以找到整个系统树所需的变更总数，具有最小总数的系统树就是最简约的。为说明这一方法，我们讨论 Fitch(1971)所给的例子。有 6 个物种 A～F 的序列可以利用，并且在某一特定位置，它们分别具有碱基 C、T、G、T、A、A。存在许多可能的拓扑结构，其中之一如图 5.11 所示。从离现

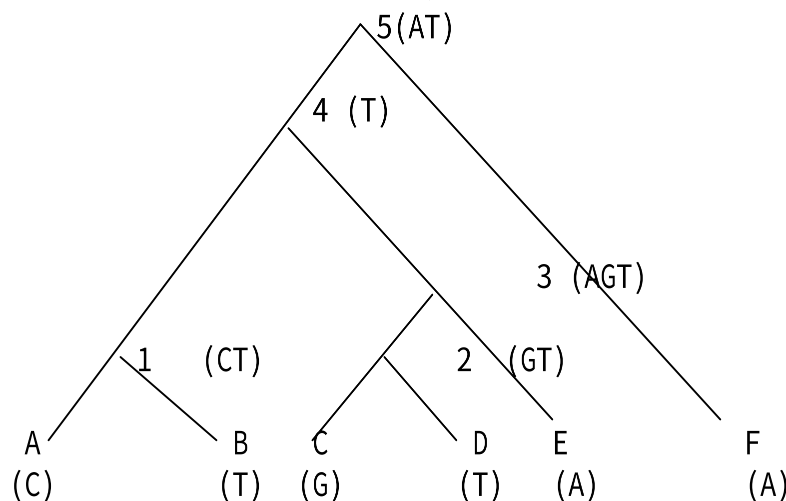


图 5.11　在 6 条序列的一个位点上寻找最简约树的过程

存序列最近的节点开始，依次考虑节点 1～5 中的每一个。在每一节点，写出两后裔序列的“简约式”。这一计算(这里记为◇)是一个集运算，如果交集不是空的，则定义此运算为两个集的交；如果交集是空的，则定义为两个集的并。对于不同的集(序列)X、Y、Z，并和交的集合运算可以与简约运算对比如下：

交　[X、Y]∩ [X、Z] = [X]　　[X]∩[Y]=ϕ

并　[X、Y]∪ [X、Z] = [X]　　[X]∪[Y]=[X，Y]

约减　[X、Y]◇[X，Z]＝[X]　　[X]◇[Y]=[X，Y]

如果两个序列在某位置具有相同碱基，则当它们的共同祖先也具有该碱基时就产生最小的变更数。如果它们具有不同的碱基，最小变更数则要求它们的祖先具有这两个碱基的其中之一。在图 5.11 中，节点 1 和 2 分别为(CT)和(GT)，意味着所列两个碱基之一将给出最小的变更数。对于节点 3 有 3 种可能性，但对于节点 4 只有 1 种可能性，节点 5 有 2 种可能性。如果节点 1～5 都具有碱基 T，则这一拓扑方法所得最小变更数为 4。但正如 Nei(1987)指出的，如果每节都有碱基 A，则产生相同的最小变更数。同时存在另外 9 种产生最小变更数的可能性，即 5 个节点具有碱基 TTTTA、TAAAA、CAAAA、AGAAA、ATAAA、CGAAA、CTAAA、TTAAA 或 TGAAA 之一者。

重复进行上述过程得到其它的拓扑结构，需要最小变更数的拓扑结构可看成为最后的系统树。对于最大化的简约，只需考虑那些信息位点(Informative

site)。对于 DNA 序列，信息位点是指那些至少存在 2 个不同的碱基且每个不同碱基至少出现两次的位点。只有一个碱基且只在一个序列中出现的位点不属于信息位点，因为那种独特的碱基位点是由于在直接通向它所在序列的分枝上发生单个碱基变更所引起的。这种碱基变更可与任何拓扑结构相容。以表 5.6 为例，只有位点 5、7、9 为信息位点。

表 5.6 信息位点列举(以 4 条序列共 9 个位点为例)

序列	位点								
	1	2	3	4	5	6	7	8	9
1	A	A	G	A	G	T	G	C	A
2	A	G	C	C	G	T	G	C	A
3	A	G	A	T	A	T	C	C	G
4	A	G	A	G	A	T	C	C	G

对于图 5.4 中的线粒体序列，存在 5 个信息位点：25、39、44、47、54。图 5.12 显示了根据这 5 个位点所得到的简约系统树。象构建其它可能系统树那样，它有 6 个碱基变更。尽管获得了与距离矩阵法找到的系统树相同的拓扑结构，但非常有限的资料已产生了某些惊人的效果。图 5.8 中节点 E 的 G 之间的分枝短于节点 G 的 F 之间的，而在信息位点间，前一分枝上有 3 个碱基变更，而在后一分枝上未发生碱基变更。

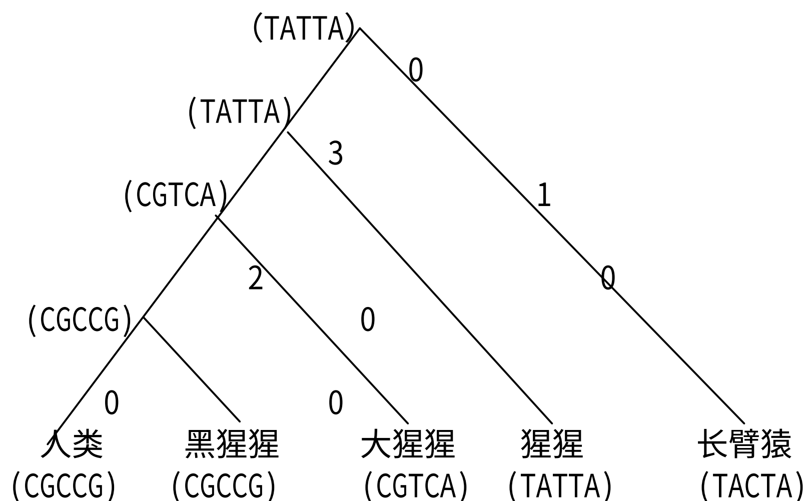


图 5.12 图 5.4 线粒体序列资料的最简约系统树
(数字为节点间的碱基变更数)

Felsenstein(1983)已批评了约减法，因为该法不是以统计原理为基础。Felsenstein 指出，在试图使进化事件的次数最小时，简约法隐含地假定这类事件是不可能的。如果在进化时间范围内碱基变更的量较小，则简约法是很合理的，但对于存在大量变更的情形，随着所用资料的增加，简约法可能给出实际上更为错误的系统树(Felsenstein,1978)。

第四节 似然法

一. DNA 序列的似然模型

构建系统树的似然法试图避免其它方法的局限性，尽管它需要的计算量大得惊人。与距离矩阵法不同，似然法试图充分有效地利用所有资料而不是将资料简缩为距离的集合。它们与简约法的不同之处在于其进化概率模型采用了标准的统计方法(Felsenstein，1981)。

当考虑实施最大似然法时，该方法先假定系统树的形式，然后选择分枝长度以使产生特定系统树的资料的似然值最大化。通过比较不同系统树的似然函数值，将具有最大似然值的系统树看作最佳估计。一个直接的问题是随着OUT的增加，系统树的数目迅速增加。当树端具有n个OUT时，无根分歧树(在每一内部树节上连接着两个分枝的树)的数目为$(2n-5)!/[(n-3)!2^{n-3}]$。当n=3、4、6、8 和 10 时，该数分别为 1、3、105、10395、2027025。具有n个树端的有根树数目与具有n+1 个树端的无根树数目相同(Felsenstein，1978)。实际应用时，只研究所有系统树的一个亚集。

对于DNA序列资料，似然法依据的模型规定了在特定时间内由于突变使一个序列变更为另一序列的概率。尽管DNA序列中的毗邻碱基不是独立的，但是模型的确假定了不同位点上进化的独立性，从而某系统树上一组序列的概率就是序列上每一位点概率的乘积。在任何单一位点，在经过时间T后，碱基i将变更为碱基j的概率为$P_{ij}(T)$。设定对于碱基A、C、G、T，下标i、j的值为 1、2、3、4。

最为简单的碱基替换突变模型假定突变率为常数。当碱基突变时，它以常数π_i的突变率变更为i型碱基。这包括了一个碱基突变为与之相同的类型，尽管这种类型的替代是观察不到的。当单位时间(世代)的碱基替换率为u时，则经过T世代后某一位点不发生突变的概率为$(1-u)^T$，因此突变概率p为：

$$P = 1-(1-u)^T \approx 1-e^{uT} \qquad (5.10)$$

经过时间 T 后由碱基 i 变更为碱基 j 的概率可写为(Felsenstein，1981)：

$$P_{ii}(T) = (1-p) + p\pi_i$$

$$P_{ij}(T) = p\pi_j \quad , \qquad (j \neq i) \qquad (5.11)$$

当设定所有π_i均为 1/4 时，这就是Jukes-Cantor突变模型，但有关突变率的解释略有不同。本模型中突变率u是对所有碱基替换而言，且u等于 4/3 乘以Jukes-Cantor模型中的可检测替换率μ。

注意到概率只涉及突变率和时间的乘积，采用这里讨论的方法无法对二者作分别估计。因此，我们只讨论乘积$\nu = uT$，即沿系统树分枝碱基替换的期望数。如果树的所有分枝以相同的速率发生碱基替换，则分枝长度将显示出树上每对树节间的相对时间。

似然法假定了系统树的结构。现存的序列形成系统树的树端，而其它树节的序列均不知道。有关系统树资料的似然值必须考虑这些未知序列的所有可能性。

在这里所描述的一个参数突变模型下，预期 4 种碱基变具有相等频率，结果对于i=1、2、3、4，π_i设定为 0.25。另一可能的方式是利用从构建系统树的序列得到的碱基平均突变率。

二．两条序列系统树

具有两个序列的一个有根系统树如图 5.13 所示。对于这个序列的第j个核苷酸位置，观测到的碱基为S_1、S_2。设在未知祖先序列中该位点碱基为k。将所有可能为k碱基的概率相加，则该位点似然值*L(j)*为：

$$L(j)=\sum_{k=1}^{4}\pi_k P_{ks_1}(v_1)P_{ks_2}(v_2) \tag{5.12}$$

对于所有 m 个位点，似然值为：

$$L=\prod_{j=1}^{m}L(j) \tag{5.13}$$

该似然值是两个未知分枝长度v_1、v_2的函数。

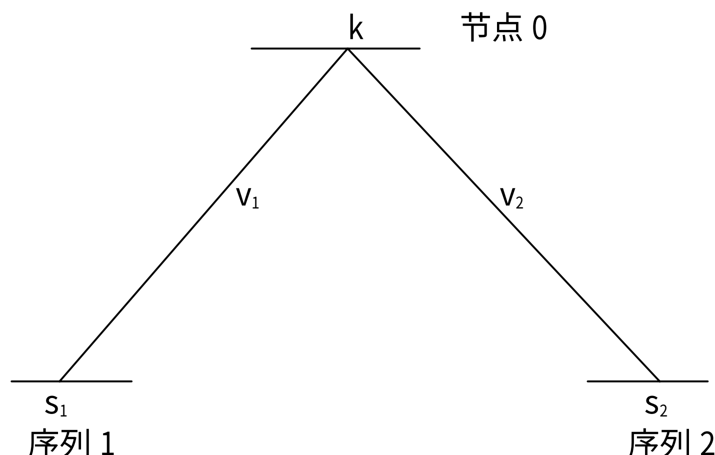


图 5.13　两个序列的有根树状图
（在j位点，两个序列具有碱基s_1和s_2和相应节点具有碱基k）

由于只存在一组从序列 1 到序列 2 的可观测的转换，因而内部节点 0 不能唯一定位。可以从 Felsenstein(1981)的“滑轮原理”来证实这一点。例如，在 j 位点序列 1 具有碱基 A，序列 2 具有碱基 C，考虑用似然函数显示该位点内部节点的 4 种碱基之和：

$$\begin{aligned}L(j)&=\pi_A P_{AA}(v_1)P_{AC}(v_2)+\pi_c P_{cA}(v_1)P_{CC}(v_2)\\&\quad+\pi_G P_{GA}(v_1)P_{GC}(v_2)+\pi_T P_{TA}(v_1)P_{TC}(v_2)\\&=\pi_A[(1-p_1)+p_1\pi_A]p_2\pi_C+\pi_C p_1\pi_A[(1-p_2)+p_2\pi_C]\\&\quad+\pi_G p_1\pi_A p_2\pi_C+\pi_T p_1\pi_A p_2\pi_C\\&=\pi_A(p_1+p_2-p_1p_2)\pi_C\\&=\pi_A p_{12}\pi_C\end{aligned} \tag{5.14}$$

换言之，涉及突变概率为p_1和p_2的两条通径(由k到A和由k到C)的似然值，与涉及概率为p_{12}的一条通径(A到C)的似然值相同。注意到

$$p_{12} = p_1 + p_2 - p_1 p_2 = 1 - e^{-(v_1+v_2)} \qquad (5.15)$$

因而图 5.13 系统树的似然值只取决于两个物种 1 和 2 间总的分枝长度(v_1+v_2)，而与节点 0 的位置无关。不可能分别估计v_1和v_2，因而系统树简缩成两个序列间的单个分枝。换言之，可估计得到的系统树是无根的。

当 4 种碱基的概率相等时，即π_i=1/4(i=1、2、3、4)，则该一分枝系统树的似然值简缩为：

$$L = \left(\frac{4-3p}{64}\right)^s \left(\frac{p}{64}\right)^{m-s} \qquad (5.16)$$

其中 p 是该分枝的突变概率，且两个序列的 m 个位点中有 s 个具有相同的碱基。将似然值最大化，得到

$$\hat{p} = \frac{4(m-s)}{3m} \qquad (5.17)$$

分枝长度的最大似然估计值为

$$\hat{v} = \ln\left(\frac{3}{4\tilde{q}-1}\right) \qquad (5.18)$$

其中

$$\tilde{q} = \frac{s}{m}$$

回顾一下，u 与 Jukes-Cantor 模型中的 4μ/3 相对应，且两序列间的时间 T 在那个模型中写作 $2t$(从每一序列到祖先序列的时间的两倍)。这些关系表明，分枝长度也可以从两个序列间的 Jukes-Cantor 距离 K 得到：

$$v = uT = \ln\left(\frac{3}{4q-1}\right)$$

$$K = 2\mu t = \frac{3}{4}\ln\left(\frac{3}{4q-1}\right) \qquad (5.19)$$

长度 v 是所有碱基替换的期望数，而长度 K 是指可检测到的替换，且 v=4k/3。

三．三条及多条序列系统树

对于三个序列则存在三种有根系统树形式，其中之一如图 5.14 所示。除了三个可观测的序列外，在节点 0 与 4 还有未定的序列，且有 4 个分枝长度有待确定。可依次考虑三种树状图，给出最大似然值的就是估计得到的系统树。但事实上，没有必要这样做，因为三种树状图具有相同的似然函数。

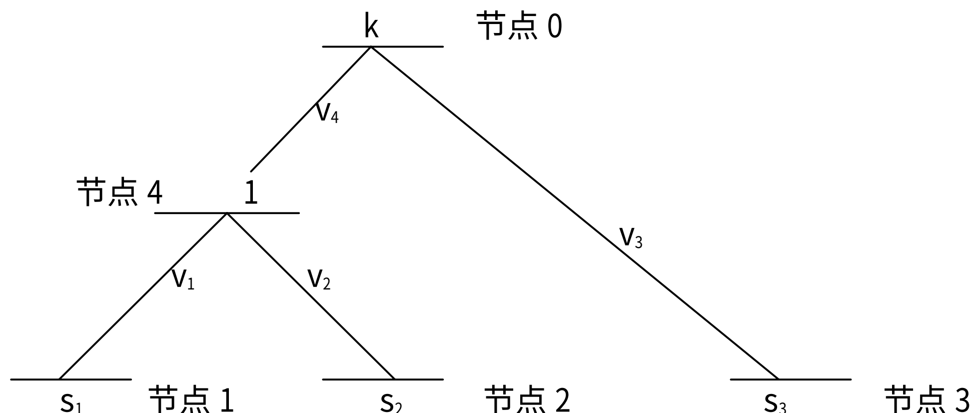


图 5.14　　三个序列的一种有根系统树形式
（在位点j，三个序列具有碱基s_1、s_2、s_3，节点 0 和 4 具有碱基k和l）

对于图 5.14 所示的排列方式，位点 j 的似然值可以用节点 4 的碱基 l、节点 0 的碱基 k 表示如下：

$$L(j)=\sum_k\sum_l \pi_k P_{kl}(v_4)P_{ks_3}(v_3)P_{ls_1}(v_1)P_{ls_2}(v_2) \qquad (5.20)$$

如果节点 0 移动到节点 3 和 4 之间的任何位置，则Felsenstein滑轮原理的应用不会改变该似然值。似然值只取决于总距离v_3+v_4。如果使节点 0 和 4 叠合，则似然值可写作：

$$L(j)=\sum_k \pi_k P_{ks_1}(v_1)P_{ks_2}(v_2)P_{ks_3}(v_3) \qquad (5.21)$$

无法唯一地确定接点 0 的位置，且对于三个序列只有图 5.15 中星状系统树需要考虑。

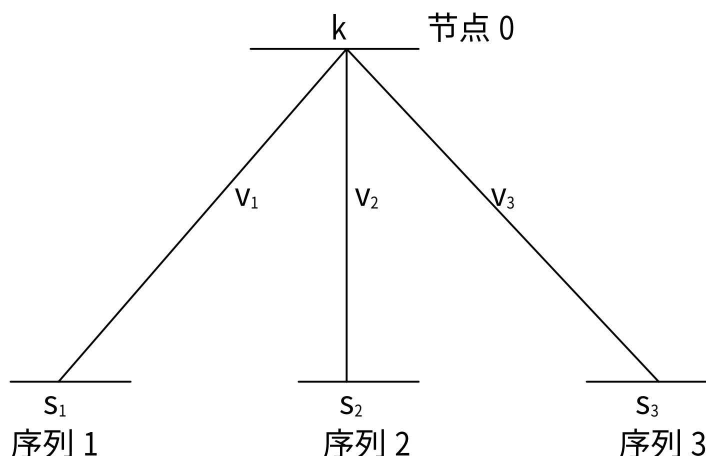


图 5.15　三个序列的星状系统树
(三个序列 1、2、3 来自于同一祖先序列 0)

在相等碱基频率的假定下，由于存在三个未知的分枝长度且有三个成对的 Jukes-Cantor 距离可供利用，所以利用 Bailey 法可从下列等式得到最大似然估

计：

$$\hat{v}_1 + \hat{v}_2 = K_{12}$$

$$\hat{v}_1 + \hat{v}_3 = K_{13}$$

$$\hat{v}_2 + \hat{v}_3 = K_{23}$$

估值为

$$\hat{v}_1 = \frac{1}{2}(K_{12} + K_{13} - K_{23})$$

$$\hat{v}_2 = \frac{1}{2}(K_{12} + K_{23} - K_{13})$$

$$\hat{v}_3 = \frac{1}{2}(K_{13} + K_{23} - K_{12})$$

实际序列并非具有相等的碱基频率，因而Jukes-Cantor距离不会使似然值最大，但它们的确为迭代法提供了很好的初始值。Newton-Raphson迭代法为找到最大似然值的数值解提供了直接的方法，且从寻求$p_i=1-e^{-v_i}$的估值来看，这一方法在描述上是最为简单的。

表 5.7 给出了图 5.4 中人类(1)、大猩猩(2)、长臂猿(3)线粒体序列收敛过程的例子。三个序列间的平均碱基频率用作模型中的概率项π_i。

表 5.7 图 5.4 中人类、大猩猩和长臂猿线粒体序列非约束型最大似然树分枝长度的连续迭代

迭代	v_1	v_2	v_3
初始值	0.0423	0.0174	0.2215
1	0.0420	0.0196	0.2230
2	0.0420	0.0199	0.2299
3	0.0420	0.0199	0.2299
标准差	0.0297	0.0218	0.0600

用几个序列作为树端来构建系统树时，可采用以上所述的一般方法。先指定一种系统树，然后对来自该系统树似然函数的方程进行 Newton-Raphson 迭代来估计分枝长度。在理论上，应研究所有可能的系统树来寻找具有最大似然值的系统树。Fukami 和 Tateno(1989)证实至多存在一组对于 L 给出平稳值的分枝长度，且这组分枝长度提供了所需的最大似然估计。将这一方法应用于图 5.4 所列的 5 种线粒体序列，获得了图 5.16 所示的无根树状图。

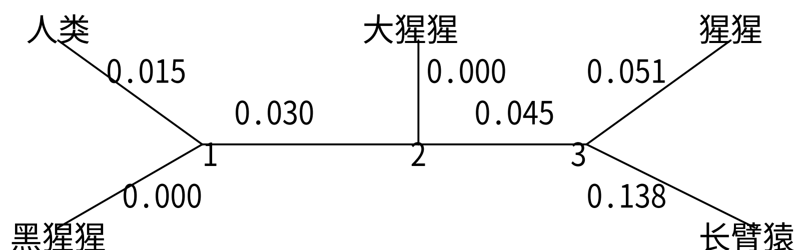


图 5.16 利用 Felsenstein 的 PHYLIP 软件构建的图 5.4 线粒体序列资料的最大似然树

四．对系统树 Bootstrap 抽样

在任一特定的树状拓扑结构内,已知最大似然值提供了分枝长度的一致估计值，这意味着随着资料量的增加，估计值逐渐接近真值。但是，与所有拓扑结构相比,具有最大似然值的系统树特性是怎样的？在何种意义上它可以认为估计了真实的系统树？尽管这是一个难以解决的理论问题,但在实际上可采用数值重复抽样来获得经验性的证据。

Felsenstein(1985)建议在所研究序列的各位点进行 Bootstrap 抽样。当序列长度为 m 时，Bootstrap 样本就包括从原始 m 个位点进行有返回抽样所得每一序列在 m 个位点的那些碱基。每一 Bootstrap 样本象原始资料一样进行相同的似然估计。对所有 Bootstrap 样本范围内应注意单源(monophyletic)物种的集合。如果发现一组物种它与 95%的 Bootstrap 系统树一起出现，则可以认为这组物种在 5%显著水平上是单源的。还有一个有用的概念，即由“多数规则”(majority rule)建立一致树(consensus tree)(Margush 和 McMorris，1981),它由在 Bootstrap 样本所得的大多数系统树中出现的那些物种所组成。在系统发育分析中获得不同的系统树时，往往需要将这些系统树组合成一致树。

第六章 蛋白质的功能域、结构及其药物设计

随着人类基因组全序列测定的完成，预示着基因组研究从结构基因组(Structural Genomics)进入了功能基因组(Functional Genomics)研究时代。研究基因组功能当然首先要研究基因表达的模式。当前研究这一问题可以基于核酸技术，也可以基于蛋白质技术，即直接研究基因的表达产物。测定一个有机体的基因组所表达的全部蛋白质的设想是由 Williams 于 1994 年正式提出的，而“蛋白质组”(proteome)一词是 Wilkins 于 1995 年首次提出。蛋白质组是指由一个细胞或组织的基因组所表达的全部相应的蛋白质。蛋白质组与基因组相对应，均是一个整体概念，但是两者又有根本的不同：一个有机体只有一个确定的基因组，组成该有机体的所有不同细胞都共享有一个基因组；但是，基因组内各个基因表达的条件、时间和部位等不同，因而它们的表达产物(蛋白质)也随条件、时间和部位的不同而有所不同。因此，蛋白质组又是一个动态的概念。由于以上原因，再加上由于基因剪接，蛋白质翻译后修饰和蛋白质剪接，基因遗传信息的表达规律更趋复杂，不再是经典的一个基因一个蛋白的对应关系，而是一个基因可以表达的蛋白质数目大于一。由此可见，蛋白质组研究是一项复杂而艰巨的任务。

蛋白质结构与功能的研究已有相当长的历史，由于其复杂性，对其结构与功能的预测不论是方法论还是基础理论方面均较复杂。统计学方法曾被成功地应用于蛋白质二级结构预测中，如Chou和Fasman提出的经验参数法便是最突出的例子。该方法统计分析了各种氨基酸的二级结构分布特征，得出相应参数(P_a,$P_β$和P_t)并用于预测。本章将简要介绍蛋白质结构与功能预测的生物信息学途径。

第一节 蛋白质功能预测

一、根据序列预测功能的一般过程

如果序列重叠群(contig)包含有蛋白质编码区，则接下来的分析任务是确定表达产物——蛋白质的功能。蛋白质的许多特性可直接从序列上分析获得，如疏水性，它可以用于预测序列是否跨膜螺旋(transmenbrane helix)或是前导序列(leader sequence)。但是，总的来说，我们根据序列预测蛋白质功能的唯一方法是通过数据库搜寻，比较该蛋白是否与已知功能的蛋白质相似。有 2 条主要途径可以进行上述的比较分析：

①比较未知蛋白序列与已知蛋白质序列的相似性；

②查找未知蛋白中是否包含与特定蛋白质家族或功能域有关的亚序列或保守区段。

图 6.1 给出了根据序列预测蛋白质功能的大致过程。由于涉及数条技术路线，所得出的分析结果并不会总是相一致。一般来说，数据库相似性搜索获得的结果最为可靠，而来自 PROSITE 的结果相对不可靠。

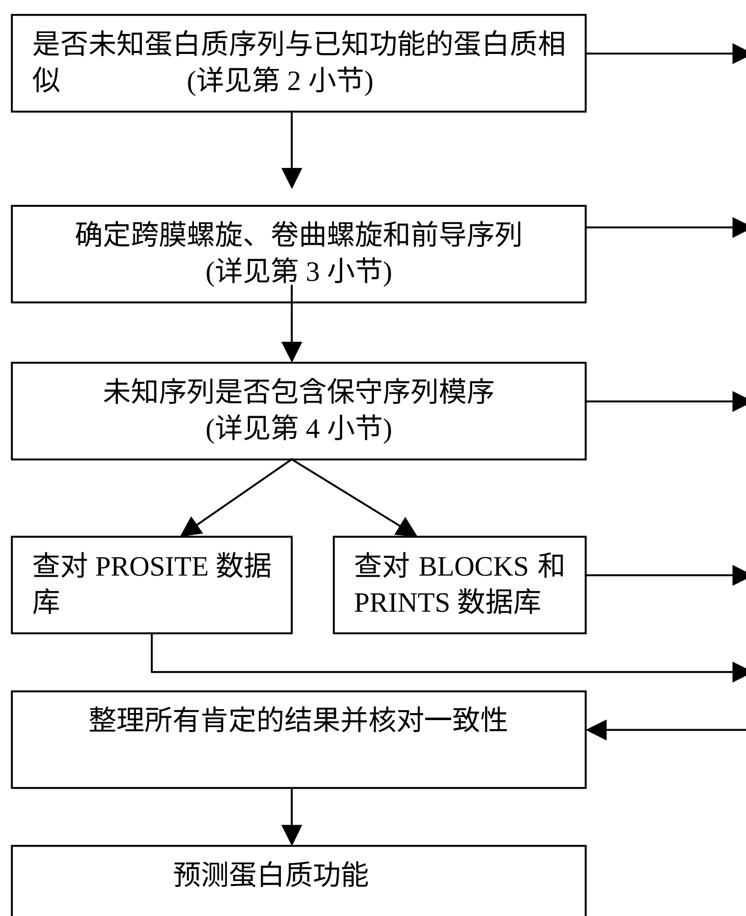


图 6.1 根据序列预测蛋白质功能的技术路线

二、通过比对数据库相似序列确定功能

具有相似序列的蛋白质具有相似的功能。因此，最可靠的确定蛋白质功能的方法是进行数据库的相似性搜索。具体的搜索方法可参见第三章，但应记住，一个显著的匹配应至少有 25%的相同序列和超过 80 个氨基酸的区段。

已有不少种类的数据库搜索工具，它们或者搜索速度慢，但灵敏；或者快速，但不灵敏。快速搜索工具(如BLASTP)很容易发现匹配良好的序列，所以没有必要再运行更花时的工具(如FASTA、BLITZ)；只有在诸如BLASTP不能发现显著的匹配序列时，这些工具才被使用。所以，一般的策略是首先进行BLAST检索，如果不能提供相关结果，运行FASTA；如果FASTA也不能得到有关蛋白质功能的线索，最后可选用完全根据Smith-Waterman算法设计的搜索程序，例如BLITZ(www.ebi.ac.uk/searches/blitz.html)。BLITZ不做近似估计(BLAST和FASTA根据Smith-Waterman算法做近似估计)，所以很花时，但非常灵敏。通常诸如BLITZ的程序能够发现超过几百个残基但序列相同比率低于 20～25%的匹配，这些匹配可能达到显著，但会被那些应用近似估计的程序错过。

还应注意计分矩阵(scoring matrix)的重要性。选用不同的计分矩阵有不少重要原因：首先，选用的矩阵必须与匹配水平相一致，例如，PAM250 应用于远距离匹配(<25%相同比率)，PAM40 应用于不很相近的蛋白质序列，而 BLOSUM62 是一个通用矩阵；第二，使用不同矩阵，可以发现始终出现的匹配序列，这是一条减少误差的办法。

除了选用不同的计分矩阵，同样可以考虑选用不同的数据库。通常可以使用的数据库是无冗余蛋白序列数据库 SWISS-PROT 和 PDB。其它一些数据库也可以试试，如可用 BLASTP 搜索复合蛋白质序列库 OWL (www.biochem.ucl.ac.uk/bsm/dbbrowser/OWL/owl_blast.html)。

二、序列特性：疏水性、跨膜螺旋等

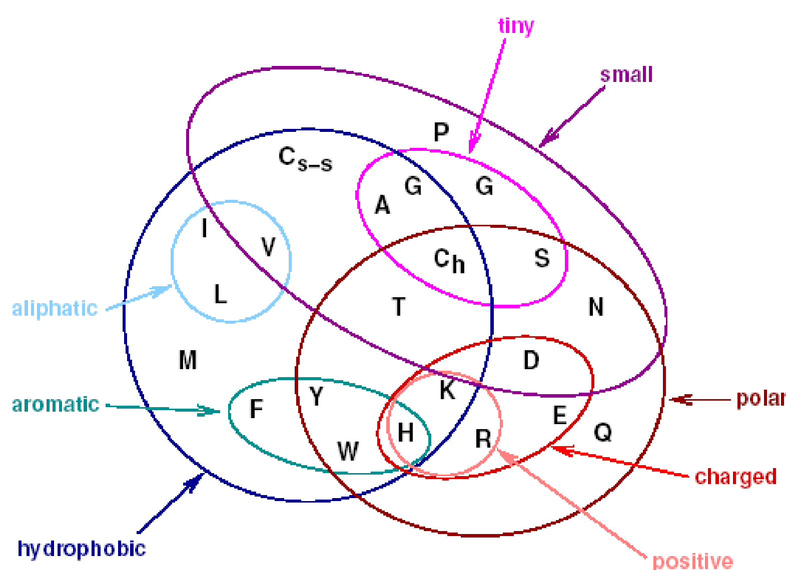
许多功能可直接从蛋白质序列预测出来。例如，疏水性信息可被用于跨膜螺旋的预测。还有不少小的模序(motif)是细胞用于特定细胞区室(cell compartment)蛋白质的定向。网上有大量数据资源帮助我们利用这些特性预测蛋白质功能。

疏水性信息可用 ExPASy(http://expasy.hcuge.ch/egibin/protscal.pl)的 ProtScale 程序创建并演示。这是一个很有用的工具，它能计算超过 50 种蛋白质的特性。程序的输入即可通过输入框将序列粘贴进去，也可输入 SWISS-PROT 的记录号。仅一项需要额外设定的参数是输入框的宽度，该参数将指示系统每次运行计算和显示的残基数，其缺省值为 9。如果想考虑跨膜螺旋特性，该参数设置应为 20，因为一个跨膜螺旋通常有 20 个氨基酸长度。图 6.2 是 ProtScal 程序的一个典型结果显示格式。

图 171 图 16.2

有多种方法可以预测序列的跨膜螺旋。最简单的方法是通过查找包含有 20 个疏水残基的区段，一些更复杂、更准确的算法不仅可以预测跨膜螺旋的位置，还能确定其在膜上的方向。这些方法都依赖于一系列已知跨膜螺旋特性的研究结果。TMbase 是一个自然发生的跨膜螺旋数据库 (http://ulrec3.unil.ch/tmbase/TMBASE_doc.html)。相关的一些程序：TMPRED (http://ulrec3.unil.ch/software/TMPRED-form.html)、PHDhtm (www.embl_heidelberg.de/services/sander/predictprotein/predictprotein.html)、TMAP (http://www.embl-heidelberg.de/tmap/tmap/tmap_sin.html)和 MEMSAT (ftp.biochem.ucl.ac.uk)。这些程序将使用了不同的统计模型，总体上，预测准确率在 80～95%左右。跨膜螺旋是可以根据序列数据比较准确预测的蛋白质特性之一。

预测前导序列或特殊区室靶蛋白信号的程序：SignalP (http://www.cbs.dtu.dk/services/SignalP) 和 PSORT (http://psort.nibbac.jp/form.html)。另一个可从序列中确定的功能模序是卷曲(coil)螺旋。在这一结构中，二个螺旋由于疏水作用而缠绕在一起形成非常稳定的结构。相关的 2 个程序：COILS (http://ulrec3.unil.ch/software/COILS_form.html) 和 Paircoil (http://ostrich.lcs.mit.edu/cgi-bin/score)。



四、通过比对模序数据库等确定功能

经常会出现这样的情况：通过列线，未知蛋白质序列与数据库内已知功能的序列均相差较大，找不到可靠的匹配结果，相反，也许会发现与某一不知功能的序列相匹配。对于这一情况，仍然可以用生物信息学工具进行一些分析。

蛋白质不同区段的进化速率不同：蛋白质的一些部分必须保持一定的残基模式以保持蛋白质的功能，通过确定这些保守区域，有可能为蛋白质功能提供线索。例如，有许多短序列可以识别蛋白质活性位点或结合区域。整联蛋白(integrin)受体识别 RGD 或 LDV 配体模序(motif)，如果未知序列中包含有 RGD 模序，则可推测未知序列的一个功能可能是结合整联蛋白。这样的推测并不是说该蛋白质序列一定会结合整联蛋白(许多含有 RGD 的蛋白质并不结合整联蛋白)，但它的确为我们提供了一个可供试验的假设。还有些例子是保守序列位于酶活性位点、转录后修饰位点、协作因子结合位点或蛋白质分类信号等，不少有关这些保守模式(pattern)的生物信息学资源已经建立起来，并已用于在序列的搜索比对。

主要有二种方法可用于序列模序的查找。一种方法是查找匹配的一致(consensus)序列或模序。该技术的优点是快捷，模序数据库庞大且不断被扩充；缺点是有时不灵敏，因为只有与一致序列或模序完全匹配才会被列出，而近乎匹配的都将被忽略。这将使你进行更复杂的分析时受到严重限制。这时，第二种方法，一种更精细的序列分布型(profile)方法将发生作用。原则上，分布型搜索的是保守序列(不只是一致序列)，这样可以更灵敏地找出那些相关性较远的序列。但是分布型和分布型数据库的创建并非易事，它需要大量的计算和人力，因此，分布型数据库的记录数并没有模序数据库多。在实际分析时，应同时对这二种类型的数据库都进行搜索，其中在一个数据库中显著的匹配可能在另一个数据库中

被完全错过，反之亦然。

最知名的模序数据库是PROSITE(http://expasy.hcuge.ch/sprot/prosite.html)。PROSITE记录的典型形式（以酪蛋白激酶II磷酸化位点的一致序列为例）：[ST]-x(2)-[DE]，即一个丝氨酸(S)或酪氨酸(T)紧跟任意 2 个残基，然后再是一个D或E。另外记录中包含了位点其它一些重要信息，如位点的作用、在何处被发现等。

分布型(profile)数据库主要有BLOCKS(http://www.blocks.fhcrc.org/blocks/)、PRINTS(http://www.biochem.ucl.ac.uk/bsm/dbbrowers/PRINTS/)和ProDom(http://protein.toulouse.inra.fr/prodom/prodom.html)。正如其它生物信息学资源一样，这些数据库总是在规模和质量之间寻求平衡。对于分布型数据库的质量来说，还包括多序列列线产生的分布型。记录数最多的数据库是依赖于自动列线程序，得到的结果有时并非是最佳结果；而记录数少的数据库一般花很多时间用于分析，人工核对列线结果，力求产生高质量的结果。一般地，分析时应搜索所有的相关数据库，以保证没有任何的遗漏。BLOCKS 数据库是利用 PROSITE 数据库模序经无空位多序列列线构建而成，PRINTS 数据库(最小的数据库)的记录来自保守序列的多序列列线，而 ProDom 数据库(version33)数据则来自 9600 个蛋白功能区模序(domain motif)的列线结果。以上列出的数据库具体情况和输出结果(有时还挺复杂)等可参照各数据库的帮助说明。

第二节 蛋白质结构预测

一、蛋白质结构及其数据库

一般情况下，蛋白质的结构分为 4 个层次：

初级结构——蛋白质序列；

二级结构——α—螺旋和β—折叠片(β-sheets)模式；

三级结构——残基在空间的布局；

四级结构——蛋白质之间的互作。

近年来，另一个介于二级和三级结构之间的蛋白质结构层次——所谓蛋白质折叠(fold)已被证明非常有用。“fold”描述的是二级结构元素的混合组合方式。

根据序列或多序列列线预测蛋白质二级结构的技术已相对比较成熟(见下小节)，但三级结构的预测则相当困难。往往对于三级结构预测，只能通过与已知结构蛋白序列同源性比对来完成。已有不少相关数据库被建立起来用于蛋白质结构预测。这一方法已是目前进行三级结构预测的最准确方法(见第三小节)。但是这一方法并不总是奏效，因为大约有 80%的已知蛋白质序列找不到与之相似的已知结构的蛋白质序列。近年来，一些新方法被提出，这些方法可以不通过相似性比对来预测序列结构。

蛋白质结构数据库主要包括 PDB、NRL—3D、HSSP、SCOP 和 CATH 等，这些数据库的基本情况及网址请参阅第二章蛋白质数据库一节。

二、二级结构预测

已有大量有关根据序列预测蛋白质二级结构的文献资料，这些资料可大致分为二类：一是有关根据单一序列预测二级结构；二是有关根据多序列列线预测二级结构。

直到最近为止，二级结构预测才不被认为具有很高的随机性。大多数预测算法均是依据单一序列。即使是最著名的一些算法(如Chou-Fasman算法和GOR算法)也只有约 60%的预测准确率，而对于一些特定的结构，如那些富含 β－折叠片的结构，这些算法难以预测成功。预测失败的原因主要是单一序列所提供的信息只是残基的顺序而没有其空间分布的信息。两个方面的研究进展改变了这一状况：一是认识到多序列列线可被用于改进预测能力。多序列列线可被视为诱变遗传学试验中的自然突变状况，其对序列上单一位点变异的分析的确提供了该位点在蛋白质三级结构中的信息；二是神经网络已开始被用于根据序列预测结构。目前已有这样一个共识，即在有大量、高质量的多序列列线结果的情况下，蛋白质二级结构的预测将非常准确——通常准确率比以单一序列预测提高 10%。一些文献表明，一些程序(诸如 PHD)预测的准确率达到了目前最高水平。PHD(http://www.embl-heidelberg.de/predictprotein/predictprotein.html)提供了从二级结构预测到折叠(fold)识别等一系列功能。

三、三级结构预测

比对数据库中已知结构的序列是预测未知序列三级结构的主要方法。多种途径可进行以上这种比对。最容易是使用 BLASTP 程序比对 NRL－3D 或 SCOP 数据库中的序列。如果发现超过 100 个碱基长度且有远高于 40%序列相同率的匹配序列，则未知序列蛋白与该匹配序列蛋白将有非常相似的结构。在这种情况下，同源性建模(homology modeling)在预测该未知蛋白精细结构方面会发挥非常大的作用。在序列相同率为 25%～40%时，两条蛋白质将具有相同的折叠，但这时同源性建模将变得更加困难和不准确。

如果在比对 NRL－3D 数据库时没有发现匹配序列,接下去可试试 HSSP 数据库。这样做的一条最方便捷径是用 BLAST 或 FASTA 法搜索蛋白质序列库(如 SWISS-PROT、TREMBL 或 PIR)，然后利用诸如 SRS 等工具去检索任何超过 25%序列相同率的匹配序列，如果这些匹配序列在 HSSP 数据库中存在，则在该序列的注释(annotation)“DR”栏中将有说明(参见第三章)。如果未知蛋白质序列与某一 HSSP 数据库序列有明显大于 25%的序列相同率,则有把握地假定未知序列至少有与 HSSP 序列相同的蛋白质折叠模式。目前，NRL－3D 和 HSSP 数据库的记录数量可以保证 20%的蛋白质序列将找到已知结构的同源序列。

总的来说，同源性建模需要专业分子建模方法和分子图象资源的辅助才能进行。不妨到Swiss－Model网站(http://expasy.hcuge.ch/swissmod/SWISS-MODEL.html)看看。Swiss－Model是一个蛋白质自动建模服务器，使用者可以直接发送一条序列或使用者自己完成的列线结果给该服务器用于同源性建模。

近年蛋白质结构研究的最主要进展之一，是有关“串线”(threading)算法和折叠识别。这些使人兴奋的技术可以在不存在已知结构同源蛋白质序列的情况下，预测所有可能的蛋白质结构。“这个未知蛋白序列会是什么结构呢？”我们也可以这样问：“我已经观察了已知结构蛋白质的各种折叠方式，未知序列是否会象这些已知结构中的某一个一样折叠呢？”第一个问题涉及几十亿种可能结构的搜索，而第二个问题涉及的是少于 1000 种结构的搜索。特定的蛋白质折叠被一而再，再而三地观察到——大部分新的经晶体衍射的蛋白将会与我们已知的折叠相关，这些过程使预测的成功机率不断提高。在串联算法中，未知序列以合适的方式被“串”到一个数据库某一折叠模板，然后计算该序列的能(energy)；在该序列与数据库中所有的折叠模板均“串”好后，可以进行计分比对，决定那些匹配达到了显著。折叠的识别技术目前还不是特别可靠的技术，只有在序列相同比率在 30%～50%时，

才有可能获得准确的估计。相关程序的结果也相当粗糙，大多数情况下难以作为同源性建模研究的依据。但是它是大多数蛋白质结构预测信息唯一可利用的工具。一些相关应用程序：TOPITS(http://www.embl-heidelberg.de/predictprotein/predictprotein.html)、frsvr(http://www.mbi.ucla.edu/people/frsvr/frsvr.html)、123D(http://www_lmmb.ncifcrf.gov/~nicka/123D.html)、THREADER 和 THREADER2(http://globin.bio.warwick.ac.uk/~jones/threader.html) 和 ProFIT(http:lore.came.sbg.ac.at/Extern/software/Profit/profit.html)。

第三节　计算机辅助药物设计[1]

开发一种新药需要平均 10-12 年，筛选 1.5-2 万种化合物，3-5 亿美元。开发新药有两个瓶颈问题：疾病相关的靶标大分子的确定；具有生物活性的小分子药物的设计与发现。计算机辅助药物设计（computer-aided drug design，CADD）分为间接与直接设计，其基本原理“锁钥原理”：E. Fischer(1894)提出药物作用于体内特定部位，如同钥匙和锁的关系一样

间接药物设计

- 定量构效关系（quantitative structure-activity relationship，QSAR): Hansch(1962)和 Free & Wilson(1964)提出。不考虑化合物的空间结构，称为 2D-QSAR。
- 3D- QSAR: CoMFA(比较分子力场分析)、距离几何（distance geometry)等
- 药效基因模型法

直接药物设计

- 以药物作用对象——靶标生物大分子的三维结构为基础,研究小分子与受体的相互作用，设计出从空间形状和化学特性两方面都可以很好与靶标分子“结合口袋”相匹配的药物分子。
- 分为全新药物设计（*de novo* drug design）和分子对接（docking）或数据库搜索两种方法。

全新药物设计
- 根据“结合口袋”的几何形状和化学特征设计药物分子
- 碎片连接法：基团或原子+适当的连接片段
- 碎片生长法：从靶标分子的结合空腔一端“延伸”出药物分子

分子对接（数据库搜索）

[1]本部分内容取自罗小民等，生物信息学与药物设计，见：赵国屏等主编，生物信息学，科学出版社，2001

⌗首先建立大量（几十到上百万）的化合物的三维数据库，然后用库中的分子与靶标分子进行“对接”（docking），选出最佳构象的分子（前 50-100 个）供药理测验。
⌗ Kuntz(1982)发展了第一个 Dock 程序，这一方法取得巨大成功

设计实例：HIV-蛋白抑制剂

第七章 小 RNA 分析

内源性非蛋白质编码小 RNA (small non-protein-coding RNA, 12-24nt)广泛存在于高等和低等生物体内，通过对靶标 mRNA 直接切除或抑制其翻译在转录后水平对基因表达起调节作用。已知的小 RNA 主要分为两大类：一类是微小 RNA (miRNA, microRNA)，一类是小干扰 RNA (siRNA, small interfering RNA)。在植物和动物体内，miRNA 与 siRNA 的产生机制和作用形式均有所不同，这里主要介绍植物体内的小 RNA。miRNA 是由具有发夹结构的初级转录本 （pri-miRNA）经过一系列加工过程，包括核酸内切酶 DCL1 加工后生成，而小干扰 RNA 则是通过核酸内切酶 DCL2, DCL3 和 DCL4 对具有较好互补结构的长双链 RNA 前体进行加工形成的(Vazquez 2006)。目前发现的小干扰 RNA 种类很多，根据前体序列类型和形成机制可分为： ta-siRNAs (*trans* acting siRNAs)，nat-siRNAs (natural antisense transcript-derived siRNAs), hc-siRNA (heterochromatic siRNA), ra-siRNAs(repeat-associated siRNAs)，长茎环结构的 miRNA-like 位点（miRNA-like long hairpin）和 nat-miRNA (natural antisense miRNA)。植物中发现的小 RNA 已有相当的数量，在水稻中至今已鉴定出 451 个 miRNA (miRBase, http://microrna.sanger.ac.uk/sequences/, Release 14.0)、一个 ta-siRNA 家族 （*TAS3*）和一个 mirtron (Zhu et al. 2008)。

由于小 RNA 表达的时空特异性，导致传统的实验方法研究小 RNA 效率很低，成本较高，因此借助计算方法研究小 RNA 是一个很好的补充，大大加速了该领域的研究进程。对保守 miRNA 家族的查找，miRNA 基因簇的发现，基于 miRNA 序列特征预测特异（novel）miRNA，通过高通量测序技术（454 和 SOLEXIA）产生的小 RNA 数据 （往往超过几百或上千万条序列）处理，以及小 RNA 靶位点的预测及其进化分析，这些分析均离不开生物信息学的帮助。随着研究的深入，大量的计算方法，相关软件和小 RNA 数据库不断产生，本章将对相关内容进行介绍。

第一节 miRNA 的主要特征及计算识别

一. miRNA 的主要特征

在植物体内，miRNA基因首先通过Pol II酶转录产生一个具茎环结构的miRNA初级转录本 (pri-miRNA) (Lee et al., 2004)，然后在DCL1酶 (Dicer-like enzyme)的作用下切除茎结构的尾巴或loop结构由miRNA前体 (pre-miRNA)得到miRNA:miRNA*双链复合体 (Tang et al., 2003; Kurihara and Watanabe, 2004)。miRNA:miRNA*复合体的两个3'端均有两个碱基的错位，其碱基结合允许一定的错配数，但通常不超过4个，并且没有较大的空位或loop结构。最后双链由解旋酶切开，miRNA*降解，成熟miRNA序列结合到靶基因位点进行调节， 根据与靶位点结合的紧密程度决定了对目标mRNA切割或是抑制其表达 (Bernstein et al., 2001; Papp et al., 2003; Bartel, 2004, 图1)。

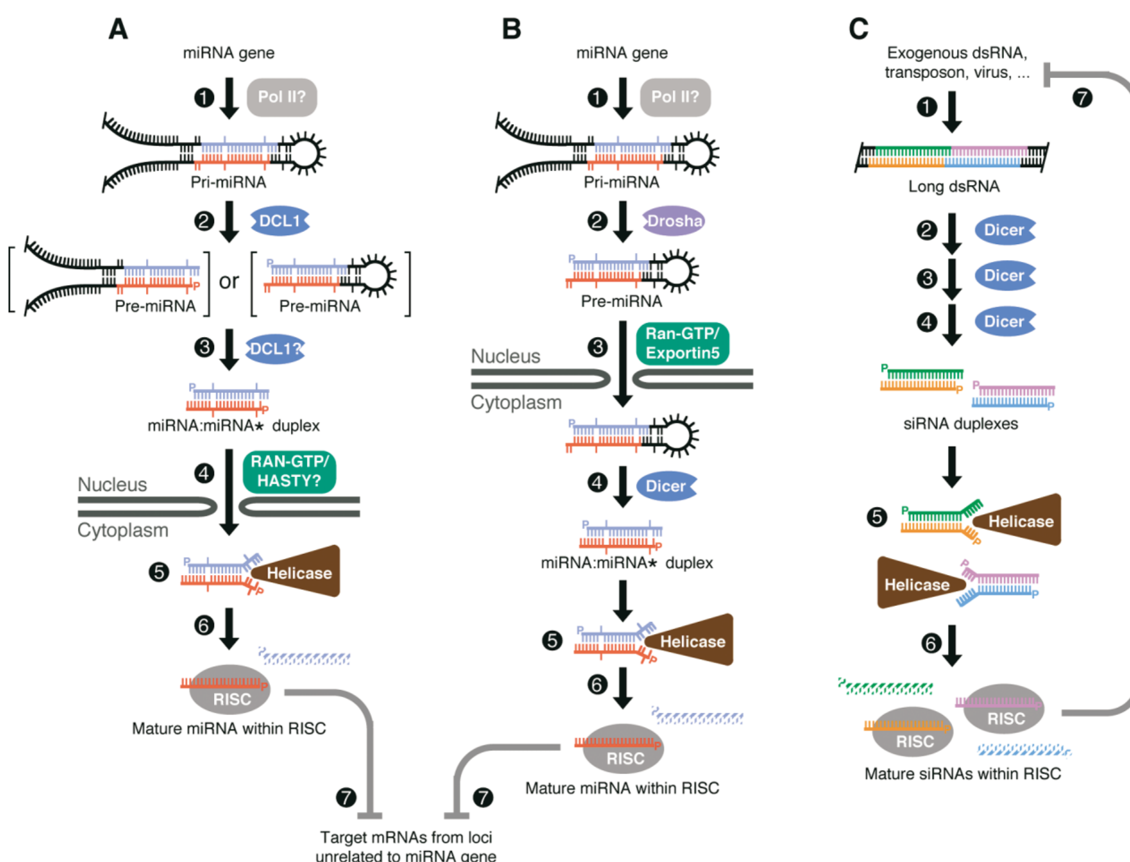


图1 miRNAs和siRNAs的产生途径 (Bartel, 2004)

(A) The biogenesis of a plant miRNA (steps 1–6; see text for details) and its hetero-silencing of loci unrelated to that from which it originated (step 7). The pre-miRNA intermediates (bracketed), thought to be very short-lived, have not been isolated in plants. The miRNA (red) is incorporated into the RISC (step 6), whereas the miRNA* (blue) is degraded (hatched segment). A monophosphate (P) marks the 5_ terminus of each fragment.

(B) The biogenesis of a metazoan miRNA (steps 1–6; see text for details) and its hetero-silencing of loci unrelated to that from which it originated (step 7).
(C) The biogenesis of animal siRNAs (steps 1–6; see text for details) and their auto-silencing of the same (or similar) loci from which they originated (step 7).

miRNA基因长度从几十到几百碱基不等（Zhang et al., 2006a,b)，但成熟miRNA序列长度一般为20-24个碱基（Ambros, 2001)，水稻中以21nt和24nt两种长度miRNA含量最丰富，这跟其选择的DCL酶有关。miRNA成簇排列的现象在动物中比较常见，在植物中目前已发现几个miRNA家族像水稻中的miR169，miR395，也在基因组上成簇排列（Jones-Rhoades and Bartel, 2004; Zhang et al., 2006a)。成簇排列的miRNA类似多顺反子结构，基因表达模式和时期均有同步性（Bartel, 2004; Altuvia et al., 2005; Baskerville and Bartel, 2005)。

基于miRNA前体的二级结构，一些研究发现miRNA前体有较低的最小折叠自由能（MFE, minimal folding free energy)，由于MFE跟序列长度相关，Zhang等（2006b）提出了最小折叠自由能指标（MFEI, minimal folding free energy index)的概念，将序列长度考虑进来，从而为不同长度miRNA前体的MEF比较提供了一个标准，并给出0.85作为miRNA区别于其他类型RNA的MFEI值，不失为一个预测miRNA的较理想指标。

$$\text{MEFI} = \frac{100 \times MEF / L}{(G + C)\%}$$

(L: the length of pre-miRNA)

目前miRBase 14.0 (http://www.mirbase.org/)版本中miRNA的记录已经超过1万条。其中很多miRNA家族均可以在至少2个物种中找到，其中miR159, miR171家族在目前miRBase收录的全部物种中均存在（Tab. 1)。这种miRNA的保守性对于在新物种中预测保守的miRNA非常有用。尽管miRNA前体在不同物种，或不同成员间的变异非常大，但成熟miRNA序列还是相当保守的，同一miRNA家族不同物种的homologs间往往只有1, 2个碱基的差异。这种便利促使了大量的查找不同物种间保守miRNA的研究（Llave et al., 2002; Reinhart et al., 2002; Bonnet et al., 2004a; Jones-Rhoades and Bartel, 2004; Sunkar and Zhu, 2004; Wang et al., 2004a; Adai et al., 2005; Sunkar et al., 2005; Zhang et al., 2005)。除了保守miRNA外，不同物种中还存在很多物种特异的miRNA (species-specific miRNA)，这类进化上比较“年轻”的miRNA无疑在特定物种的形成和发育过程中扮演着重要的作用。

表1. 植物保守miRNA家族（根据miRBase 14.0和物种多少排序）

miRNA family	No. of species	miRNA family	No. of species	miRNA family	No. of species
miR-159	17	miR-394	6	miR-1510	2
miR-171	17	miR-157	4	miR-1514	2
miR-156	16	miR-2118	4	miR-161	2
miR-166	16	miR-824	4	miR-2111	2
miR-167	16	miR-1507	3	miR-2275	2
miR-396	15	miR-2119	3	miR-413	2
miR-160	14	miR-403	3	miR-414	2
miR-399	14	miR-437	3	miR-415	2
miR-169	13	miR-444	3	miR-416	2
miR-172	13	miR-477	3	miR-417	2
miR-319	13	miR-529	3	miR-418	2
miR-408	12	miR-530	3	miR-419	2
miR-164	11	miR-535	3	miR-420	2
miR-168	11	miR-827	3	miR-426	2
miR-162	10	miR-1122	2	miR-472	2
miR-390	10	miR-1127	2	miR-479	2
miR-393	9	miR-1135	2	miR-783	2
miR-395	9	miR-1139	2	miR-821	2
miR-398	9	miR-1432	2	miR-828	2
miR-397	8	miR-1435	2	miR-845	2
miR-482	7	miR-1509	2		

miRNA通过与靶基因形成互补RNA双链来行使调节功能，这种互补性在进化过程中是保守的 (Rhoades et al., 2002; Jones-Rhoades and Bartel, 2004; Robins et al., 2005a)。互补性的强弱或者说互补碱基的多寡决定了miRNA调节的不同机制。跟靶基因有较好互补的miRNA主要通过对目标mRNA的直接切割调节mRNA的表达，相反，如果miRNA与其靶位点的错配较多，则主要通过转录后抑制的方式干扰mRNA的翻译 (Papp et al., 2003; Bartel, 2004, 图2)。植物miRNA的靶基因一大类都是转录因子 (transcriptional factor)，揭示了miRNA调节通路的复杂性。

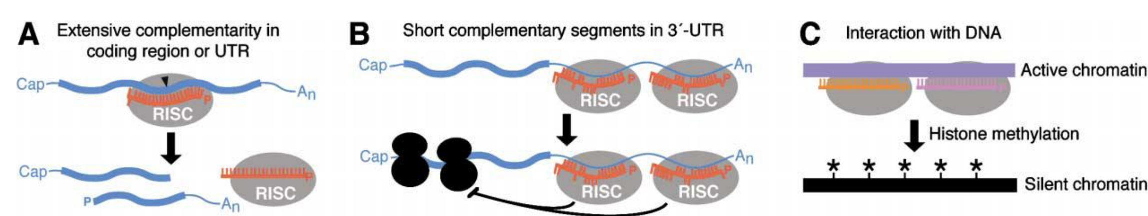


图 2 小RNA调控机制 (Bartel, 2004)

(A) Messenger RNA cleavage specified by a miRNA or siRNA. Black arrowhead indicates site of cleavage.

(B) Translational repression specified by miRNAs or siRNAs.

(C) Transcriptional silencing, thought to be specified by heterochromatic siRNAs.

二. miRNA的计算识别

通过计算方法识别miRNA基因主要基于以上提到的miRNA序列及结构上的特征，以及不同物种间的保守性。可以分为以下几类方法：

1. 同源比对

同源比对的方法主要是通过已知保守miRNA的在不同物种间的序列相似性进行同源序列搜索预测miRNA的方法。以已知miRNA序列为索引，公共DNA序列数据库中的数据作为搜索库，对于全基因组已测序或正在测序的模式生物，如rice，maize等，可利用其全基因组或大规模测序数据；对于基因组序列并未获得的物种来说，小规模的GSS (genome survey sequences)序列和EST (expressed sequence tags)序列也是很好的数据资源。尤其是EST序列，因为其本身就是表达水平的序列，故而预测的结果更加准确可信。搜索程序可以选择BLAST，如果是利用成熟miRNA序列进行搜索，因为序列较短，E值一般要高于1E-2，最小字符长度改为7 (默认13, -W 7)，但利用BLAST比对仍然会因程序本身的原因造成敏感性的降低，笔者在实际数据处理过程中曾发现对于~20nt的miRNA，2个不连续且距离较近的错配会导致错配序列3'端完全略掉联配过程，从而漏掉一个可能的结果，尽管这种情况是极少的。另外，基于轮廓的搜索软件ERPIN (http://tagc.univ-mrs.fr/erpin/)也可以用来搜索数据库中的miRNA同源基因位点。通过提交一组特定RNA的联配序列及二级结构信息，ERPIN可以搜索特定模式的RNA序列，从而获得更加准确特异的结果。同源比对方法还要注意以下几点：1) 数据处理过程中一般先通过BLASTX搜索蛋白质数据库，以排除掉编码蛋白序列，提高检索效率； 2) 往往仅找到已知miRNA的同源序列还远远不够，一般需要对候选miRNA位点周围的序列进行二级结构预测，以确定该段序列是否可能形成stem-loop结构，并需要验证miRNA的位置，及miRNA与miRNA*的互补情况；3) 在确定了可能的miRNA前体序列后，需要计算该段序列的MEF及MEFI值，一般情况下miRNA前体的MEF很小，而MEFI > 0.85，如果所有以上标准均符合，那么该位点即为候选的miRNA基因。

基于同源搜索方法开发了很多软件，包括Wang等（2005b）开发的miRAlign软件 (http://bioinfo.au.tsinghua.edu.cn/miralign/) (图3)，可以用来预测人miRNA基因的基于概率共同学习模型开发的ProMiR (cbit.snu.ac.kr/~ProMiR2/) (Nam et al.,

2005)，以及原理相似，用于植物miRNA预测的microHARVESTER (http://www-ab.informatik.uni-tuebingen.de/brisbane/tb/index.php?view=microharvester) (图4) (Dezulian et al., 2006)。

MiRAlign

Input your sequence and parameters

please inpute your sequence here(sequence length should between 50~300)

Example sequence:
ccagccugcugaagcucagagggcucugauucagaaagaucaucggauccgucugagcuuggcuggucgg

Type: animal plant

delta_len: 15

min_seq_sim: 70

MFE: - 20 kcal/mole

submit reset

图 3. miRAlign界面 (http://bioinfo.au.tsinghua.edu.cn/miralign/)

microHARVESTER on the NCBI EST database (NCBI EST est_others: all non-human and non-mouse seqs as of 27-July-2005)

Input

Enter precursor sequence(s)

Enter mature sequence(s)

[5 sequences max for one job]

Input examples

Try one of these miRNAs as your query: ATH-MIR169a ATH-MIR172a ATH-MIR390a

You might want to take a plant query from the miRNA registry.

Output examples

This is the output for the above example queries: ATH-MIR169a ATH-MIR172a ATH-MIR390a

Instructions

Find detailed instructions here.

Job Options

Job-ID job33276

Please avoid special characters in any input field. Best would be only letters and digits. Choose a unique job ID.

Reset form Submit job

图 4. microHARVESTER界面

(http://www-ab.informatik.uni-tuebingen.de/brisbane/tb/index.php?view=microharvester)

2. 基因查找

基因查找方法可以不考虑miRNA的保守性，对整个基因组进行扫描，但只适用于动物miRNA基因的预测。首先根据不同物种的全基因组联配信息确定保守的非编码区，特别是启动子区及3' UTR区 (Xie et al., 2005a)，然后设定一个窗口大小比如110nt在该区域内滑动，利用二级结构预测软件比如Mfold (http://dinamelt.bioinfo.rpi.edu/download.php)或RNAfold (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi) (图5)对每条110nt长度的序列进行二级结构预测并打分，给出候选的miRNA基因。目前有两个基于该方法的软件成功预测了动物miRNA基因。一个是miRscan (http://genes.mit.edu/mirscan/) (图6)另一个是miRseeker (Lim et al., 2003b)。Lai等（2003）在果蝇基因组中的miRNA基因鉴定

工作表明，以已知的miRNA基因做参照，miRseeker的准确度和灵敏度为75% (18/24)，但是由于两种方法都是基于一定的窗口大小对保守区域进行扫描，因此该方法对于miRNA基因序列长度变化较大的植物miRNA预测来说并不适合。

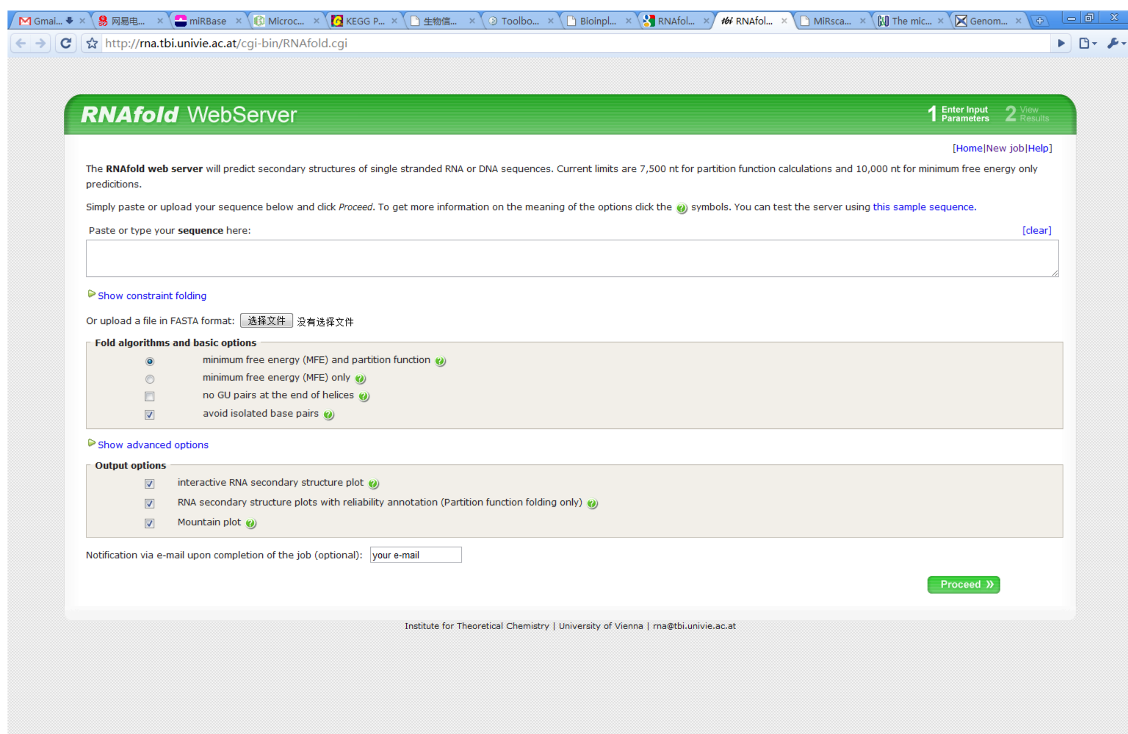


图 5. RNAfold 界面 (http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi)

以 RNAfold 为例来说明二级结构预测软件的使用。RNAfold 是 Vienna RNA Package 的一系列用于二级结构预测和计算的工具之一。作为一个开源软件，Vienna RNA package 支持 Unix/Linux/Windows 多平台的版本下载 (http://www.tbi.univie.ac.at/RNA/)，每个软件均有详细的说明文档。上图是 RNAfold 的 web server 界面。将需要做二级结构预测的序列 (RNA/DNA)粘贴到文本框中，或将保存有 fasta 格式的文件提交到 server，选择相应的参数，如果不想在线等待结果，可以提供一个 email，在程序运行完毕后会将结果的链接发到邮箱里。点击"Proceed"，最佳二级结构结果会在新窗口中显示，包括方便批量处理的“点-括号”格式结果，最小自由能值，以及图形化的结果，可保存为.eps 或.pdf 格式的文件。



图 6. MiRscan 界面 (http://genes.mit.edu/mirscan/)

3. 邻近茎环结构搜索

基于动物miRNA经常成簇存在于基因组上的特点，通过对已知miRNA附近区域进行茎环结构预测来发现成簇存在的miRNA。近期研究表明42%的人类miRNA基因和50%的果蝇miRNA基因都有成簇存在的现象 (Bartel, 2004; Altuvia et al., 2005)。由于植物miRNA成簇存在的现象比较少，只有miR169,miR395等几个家族存在成簇分布，因此该方法在植物miRNA预测方面存在局限性。

4. 基于比较基因组学的算法

基于比较基因组方法代表性研究是Jones-Rhoades和Bartel (2004)利用拟南芥和水稻全基因组鉴定在两个物种中保守的miRNA序列。作者开发了MIRcheck软件(http://web.wi.mit.edu/bartel/pub/software.html)通过计算一段序列是否存在理想的茎环结构，以及是否有20mers的短序列位于茎的位置上，然后根据其在两个物种中的保守性来查找保守的miRNA基因。Adai等 (2005)开发了findMiRNA (http://sundarlab.ucdavis.edu/mirna/)可以针对单个基因组来查找miRNA，findMiRNA主要依据miRNAs和其靶基因序列互补的保守性，然后利用二级结构预测软件对候选位点进行二级结构预测，找出有理想茎环结构的序列。需要注意的是，因为基

因组中很多类型的序列，如tRNA，逆转座子等元件均能形成发卡结构，因此在前期序列过滤和最终候选结果筛选方面要注意。

5. 基于大规模测序数据的发掘方法

从以上方法可以看出，大部分方法的理论基础都是miRNA的序列保守性，只有基因查找可以从miRNA的结构出发鉴定新的或物种特异的miRNA，但由于它是以一定长度为限制进行扫描，因而该方法对植物miRNA的预测并不适合。随着新一代测序技术如454和solexa技术的成熟和推广，大规模的基因组数据和RNA数据不断产生。针对miRNA的solexa测序每次都可以产生百万级数量的数据。在海量的数据面前仅仅通过前面介绍的传统方法显然不能满足研究的需要，如何有效的从这些海量数据中鉴定出miRNA基因变成了一个迫切而略带挑战的课题。以水稻方面的工作为例，最近发表了几篇大规模鉴定miRNA基因的文章。其中Zhu等 (2008)以发育的水稻种子为材料最终鉴定了39个新的非保守的miRNA家族；Sunkar等 (2008) 以胁迫处理的水稻幼苗为材料鉴定了23个新的miRNA。虽然采用的计算方法略有不同，但都是基于miRNA序列和结构上的保守性进行预测。

下面以Zhu等 (2008)的工作为例说明一下大规模小RNA测序的数据处理流程。基于Solexa测序的原理，测序得到的原始读序都是一端连接了接头 (adaptor)的同一长度的序列，因此首先需要过滤掉接头和一些低质量的序列，这样得到了一个从十几个碱基到二十几个碱基不等的数据库。对于已有基因组数据的物种，比如水稻、拟南芥等，可以利用序列比对工具如BLAST将测得的小RNA匹配到基因组上 (>18nt)。这样我们就得到了一个全基因组的小RNA的分布图谱。根据全基因组的注释，排除掉匹配到重复序列区域和编码区的小RNA。这样一方面我们可以用上面介绍的方法来搜索保守的miRNA基因，另外，由于已知了小RNA序列和其位置信息，我们就可以利用一些新的标准来识别新的物种特异的miRNA基因。由于miRNA在产生过程中需要形成miRNA:miRNA*复合体，首先，根据小RNA的分布寻找候选的miRNA:miRNA*复合体。标准如下：1) 两条小RNA匹配到同一染色体的同一条链，且相距不超过400nt；2) 不允许有很多其他小RNA匹配到两条序列之间的区域（特别是有另外的小RNA跟其中一条部分配对，形成“拖尾”现象)；3) 每条小RNA在全基因组的匹配位置不能太多 （不超过10处)；4) 两条smallRNA的读序数需要相差5倍以上 （根据miRNA合成原理，miRNA*在与miRNA分开后会很快降解)。两条小RNA的配对也需要符合一定的标准 (Jones-Rhoades et al. 2006)：

1) 总共不超过7个碱基 (更严格的话可以设为4个碱基)的错配；2) 不超过3个碱基的连续错配；3) 不存在一条链上超过两个碱基错配而在另一条链上没有错配碱基的对应。满足以上条件的两条小RNA序列被当做候选的miRNA:miRNA*序列。从基因组上切下包含两条互补小RNA的序列作为候选的miRNA前体序列进行二级结构预测，根据其二级结构及两条序列所处的位置判断是否为候选的miRNA基因。

以上计算方法虽然提供了一种相对方便的鉴定miRNA的手段，而且目前大部分miRNA序列都是通过计算得方法预测出来的，但由于不同的预测方法都存在或多或少的缺陷或者假阳性，所以预测得到的候选miRNA基因仍然需要通过实验方法进行验证，包括直接克隆，Northern, PCR, 5'-RACE (5' rapid amplification of cDNA ends) (Griffiths-Jones, 2004; Griffiths-Jones et al., 2006)。

三. miRNA靶基因的预测

不像动物miRNA结合靶基因的机制那么复杂，植物miRNA主要通过接近完美的互补配对结合到靶位点，从而引发对目标mRNA的直接切割。植物miRNA和靶位点的结合有如下特征：1) 一般不超过3个碱基的错配；2) 5'端前10个碱基结合很紧密，一般只允许1个碱基的错配；3) 5'端第1，11，12个碱基因为剪切功能的关系一般不允许有错配；4) 一般没有连续的错配 (>=3个)出现。动物miRNA靶基因的预测根据结合的不同特点已经开发了很多的软件，从miRanda，TargetScan，Pictar到microTar等，但由于植物miRNA识别靶位点的模式较为简单，所以植物miRNA靶位点的预测软件相对较少，其中miRU (http://bioinfo3.noble.org/miRNA/miRU.htm)是一个网络平台，整合了已知的大部分植物mRNA和gene数据，可提供候选的小RNA，在提供的植物表达数据中预测是否有靶位点（图7）。miRU有几个参数可供设置：一是阀值，即总罚分为3分，根据不同错配类型，罚分不同；二是G:U配对，一般罚0.5分，三是INDEL，一般不超过2个，四是其他类型，即错配，总共不超过3个。然后选择需要预测的靶基因数据库，即Database1，另外还有一个Database2，是预测保守miRNA靶位点提供的参照物种，可以降低预测的假阳性。另外，Zhao et al. 又在miRU的基础上开发了psRNATarget，不仅可以提供小RNA在其植物基因数据库中预测靶位点，还可以提供自己特定的基因数据(< 70Mb)检验是否存在已知的miRNA的靶基因，另外，最灵活的服务是你可以提供特定的小RNA以及特定的植物基因数据，进行完全个性化的靶基因预测，当然你的基因数据大小有一定的限制 (<70Mb)。

miRU: Plant microRNA Potential Target Finder

The program predicts plant miRNA target genes. It reports all potential sequences complementary to the query with mismatches no more than specified for each mismatch type. In addition, each mismatch is penalized according to the mismatch type and position to the miRNA. With default settings, the minimal score among all 20mers cannot exceed 3.0. This program can also be used for siRNA specificity detection. For more information about the prediction algorithm and questions about the search result, please click here.

Enter your small RNA (19-28 nt)

Score for each 20 nt 3

G:U Wobble Pairs 6

Indels 1

Other Mismatches 3

Dataset 1 Arabidopsis thaliana mRNA (from TIGR Ath1 5)

The following fields are for reducing false positives in target prediction by detecting target complementarity conservation and are optional. Select a dataset for a different organism and provide homologous miRNA from the organism, and the program reports homologous mRNA targets with conserved complementarity. If homologous miRNA is not provided, the program will not check target conservation.

Dataset 2 TIGR Rice Genome mRNA (OSA1 release 3, 12/28/2004)

Homologous miRNA

Submit Query Reset

图 7. miRU界面 (http://bioinfo3.noble.org/miRNA/miRU.htm)

patScan是另一个可以方便进行miRNA靶基因预测的软件。patScan提供了Unix/Linux/Windows版本可在 http://iubio.bio.indiana.edu/soft/molbio/pattern/下载。patScan最初的设计是用来查找基因组特定模式的序列，Rhoades et al. (2002)首先将patScan用于miRNA靶基因的预测，并评估了这种预测方式的假阳性 (Rhoades et al., 2002; 图8)。patScan的运行需要调用两个文件，一是指定搜索的pattern文件，由相应的smallRNA序列和匹配模式组成：smallRNA_sequence[4,0,0]；另一个是用来预测的基因序列文件，Fasta格式，标题按照相应的序列类型标示为“>title|CDS ..”或“>title|cDNA ..”等等。smallRNA与靶位点的匹配标准如前所述。另外，前面提到的MIRcheck和findMiRNA软件由于在预测miRNA时需要考虑miRNA和其靶位点的保守性，故而也可用来预测miRNA靶位点。

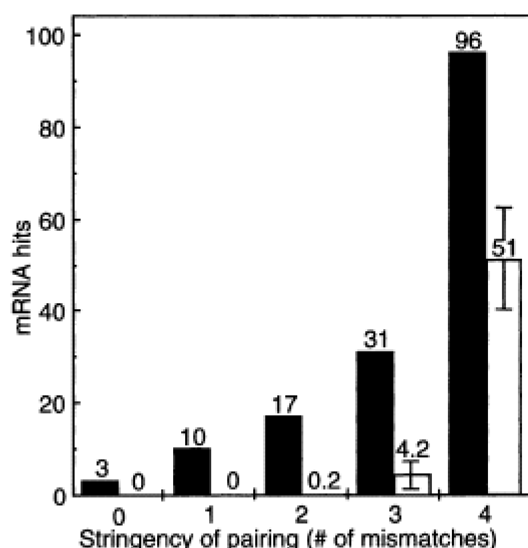


图 8. *Arabidopsis* miRNA与其mRNA的反义匹配情况 (Rhoades et al. 2002)
Annotated Arabidopsis mRNAs were searched for sites complementary to 16 Arabidopsis miRNAs with 0–4 mismatches (solid bars). Identical searches with cohorts of 16 randomized RNAs were also performed (open bars, mean values from ten cohorts; error bars, one standard deviation). Note that two hits by similar miRNAs to the same complementary site within an mRNA were counted as separate hits (Table 1).

第二节 ta-siRNA等的计算识别

一. ta-siRNA的主要特征

与miRNA不同，siRNA主要通过长的双链RNA复合体在DCL酶的切割下产生。植物体演化出几种截然不同的siRNAs，它们在产生机制和调节通路的功能方面都有所不同 (Brodersen and Voinnet, 2006; Vaucheret, 2006)。其中大部分的siRNA类型(24nt)在依赖RNA的RNA聚合酶2 (RDR2)、DCL3、PolⅣ的作用下产生，并通过AGO4引导的DNA甲基化或组蛋白修饰诱导转录沉默 (Zilberman et al., 2003, 2004; Chan et al., 2004; Xie et al., 2004; Herr et al., 2005; Kanno et al., 2005; Onodera et al., 2005; Pontier et al., 2005; Tran et al., 2005)。这一代谢通路往往跟转座子、反转座因子等重复序列相关 (Xie et al., 2004; Lu et al., 2006; Rajagopalan et al., 2006; Kasschau et al., 2007)。其他类型的siRNA主要在转录后水平起作用。对病毒RNA和转基因转录本的沉默涉及到依赖RDR6/DCL4的siRNA (21nt)或依赖DCL2的siRNA (22nt)。ta-siRNA就是通过RDR6/DCL4通路产生的。tasiRNA的形成主要是通过miRNA介导的按21nt相位排列的siRNA的剪切 (<=12 phases)。不同的TAS家族受不同的miRNA调节，TAS1和TAS2受miR173的调节，TAS3在拟南芥和水稻中保守，受miR390调节，且有5'端和3'端两个结合位点，TAS4受miR828调节。TAS基因的dsRNA前体在DCL4作用下，由相应的miRNA起始剪切，产生21nt，3'端有两个碱

基错位的双链siRNA复合体 (Dunoyer et al., 2005; Gasciolli et al., 2005; Xie et al., 2005)。不同TAS家族切割产生的siRNA数目不同，其中只有特定的一两个siRNA行使功能。根据以上特征可以通过生物信息学的方法预测tasiRNA。

二. ta-siRNA的计算识别

1. Howell算法

前面提到全基因组序列已测序的物种产生了大量的小RNA的数据，而且这些不同组织或处理下测得的小RNA可以很好的定位到全基因组上。根据一段区域(<300nt)内小RNA是否按照21nt的位移排列这一显著特征，可以找出候选的TAS基因位点。Howell 等(2007) (图9)设计了一套流程用来查找拟南芥中的候选tasiRNA，首先将定位到基因组正反链的小RNA序列合并，将来自不同链的小RNA定位位置抵消掉2个碱基，这样来自一对复合体的正反链小RNA位置可以在计算的时候累加。然后引入P值作为评价步移的参数。P值的计算如下：

$$P = \ln[(1+\sum_{i=1}^{8} k_i)^{n-2}], P > 0,$$

如果一个相位长度设为21nt，n表示在8个相位大小的窗口范围内至少有一个小RNA定位到相位上的相位循环数（即n个相位位置上有小RNA存在）；k表示在调查的这8个相位大小的窗口里面正负链合并过的起点位置刚好位于相位上的小RNA读序总和；由于指数n-2的限定，只有当至少连续三个相位上 (n>=3)都存在至少一个小RNA才能保证P为正值。由公式可以看出，P值受小RNA丰度和所处位置的双面影响。P值的计算按单碱基的步长在基因组上滑动，计算得到的P值分配给该点四个相位距离的位置。因此，可以将小RNA在基因组上的实际分布，如图9 A中READS图所示，转化为P值分布的PHASE图，具有显著高P值的位点被选为候选的phase位点。最后，根据ta-siRNA受相应miRNA调控的现象，在预测到的phase区域两端预测miRNA靶位点，如果可以找到相应的结合位点，那么这段区域可被认为是tasiRNA-like位点。

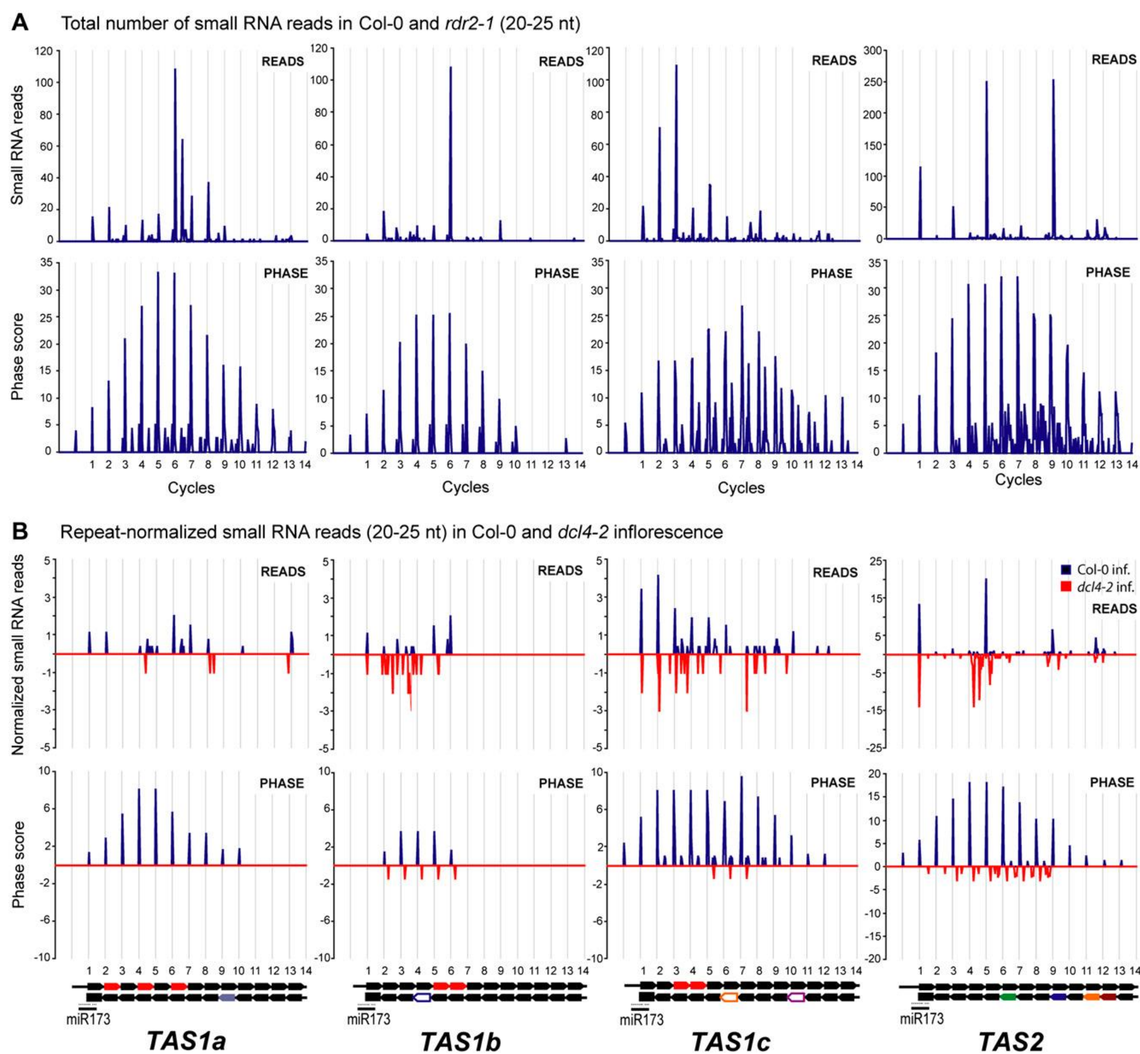


图 9. 拟南芥TAS1a, TAS1b, TAS1c和TAS2位点21nt小RNA分布及数量 (Howell et al. 2007)

2. Chen算法

与Howell的方法类似，Chen等(2007)的方法也是主要考虑tasiRNA的相位分布特征，并构建了一个 P值来查找候选的tasiRNA位点。按照21nt一个相位大小，考虑11个相位长度的一段区域，n表示位于该231bp区间的小RNA读序数；k表示位于该231bp区间相位位置上的小RNA读序数。P 值越大，表示相位（phase）结构越明显。Chen 等提供了相应的perl脚本用于计算P value，可以在其文章附件信息中找到。

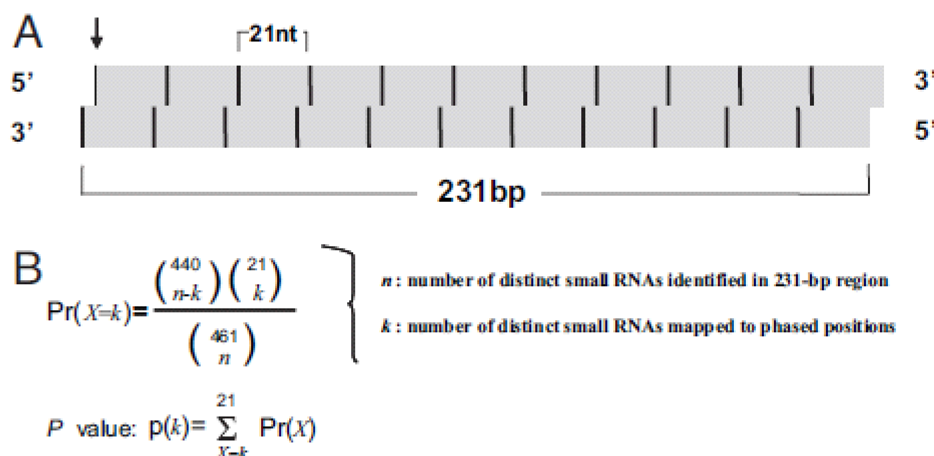


图10 TAS预测算法原理 (Chen et al. 2007)
(A) The vertical arrow indicates the start site for the small RNA used to determine the phased and nonphased positions. 21 phased sites relative to the start site are indicated as black vertical bars. Four hundred forty nonphased sites relative to the start site are indicated as gray. (B) Equation based on hypergeometric distribution for statistically evaluating the presence of phased siRNA in genomic fragment defined in A.

三. 起源于NAT的siRNA

Natural Antisense Transcripts (NAT)是指可以跟其他转录本互补形成RNA双链的编码或非编码RNA序列。根据它们在基因组上的相对位置不同，NAT可以分为两类：*cis*-NAT和*trans*-NAT。*cis*-NAT是指来自于跟有义链转录本同一个基因组座位不同染色体链的序列；*trans*-NAT是指跟它的互补序列来自于染色体上的不同位置的转录本。研究表明哺乳动物和植物中大约5%~10%的基因转录本都存在*cis*-NATs。Osato等(2003)从水稻中预测了687组NAT；Wang等(2006)从拟南芥中预测了1320个*trans*-NAT。起源于NATs位点的siRNA称为NAT-siRNA，主要介导转录后沉默。起源于NATs双链RNA复合体的小RNA称作NAT-siRNA，第一个NAT-siRNA是2005年从拟南芥中鉴定出来的，来自P5CDH和SRO5基因转录物形成的dsRNA。目前已有若干大规模鉴定NAT-siRNA的工作在拟南芥和水稻中开展，并发现了许多有意思的结果。其中包括*cis*-NAT-siRNA5'端第一个碱基的偏好性，由于AGO2和AGO4参与该类小RNA的结合，故而第一个碱基常常为腺嘌呤 (A)。对*trans*-NAT的GO分类研究表明，催化活性、信号传感器、转运蛋白活性相关的转录本占很大比例。另外，对NATs结构的功能研究表明植物基因组中的NATs结构可能对逆境胁迫方面起重要作用。

另外还有几种其他类型的siRNA，比如首先从水稻中发现的NAT-miRNA，其长约20nt。前体hpRNA序列两条链分别转录、剪接，反义链RNA产生miRNA，调节正义链mRNA的表达。NAT-miRNA既不同于普通miRNA，因为普通miRNA的前

体hpRNA无需剪接；也不同于NATs-siRNA，后者的序列多来自两条链，而nat-miRNA几乎都是由一条RNA链产生；另外，NATs -siRNA形成需要DCL2，而NAT-miRNA需要DCL1。Zhu等(2008)在水稻中发现了一类miRNA-like long hairpin位点。这类小RNA基因可以像普通miRNA那样形成长的发卡结构，但是有很大的loop环，其茎结构又跟tasi-RNA类似，在双链上有21nt的phase结构。

四. siRNA 靶基因预测

尽管siRNA有着丰富的类型，但其行使功能还是通过与靶基因位点的序列互补来实现 (图1)。因此，miRNA靶基因的预测软件也同样适用于siRNA的靶基因预测。值得注意的是，已有的研究表明，特定类型的siRNA靶基因也有着显著的区别。比如*TAS3*的靶基因是一类大的基因家族，称做激素响应因子 (ARF)。拟南芥中发现的NAT-siRNA被认为与植物的抗逆境代谢有关。

第三节 小 RNA 的进化分析

一. 小 RNA 进化研究概况

作为一类重要的调控小分子，miRNA 在大多数真核生物（Finnegan and Matzke, 2003）甚至是病毒（Sullivan et al. 2005）中通过 RNA 干扰机制调节各种代谢途径。植物中许多编码 miRNA 的基因起源于单双子叶植物分化之前（约 150 百万年前），动物中的 miRNA 编码基因也早于多细胞动物分化的时间（约 600 百万年前）。然而，目前还没有发现动植物中 miRNA 编码基因或靶基因的同源基因。这就提出一个进化上的有趣的问题：这些编码 miRNA 的基因是怎么形成的呢？

Allen 等（2004）通过对两个拟南芥特异 miRNA 家族的研究揭示了 miRNA 编码基因与其靶基因共同进化的一个可能的机制。由于 miR161/163 两个家族都是新产生的年轻 miRNA 编码基因，而且跟大多数保守的 miRNA 家族不同，miR161/163 均位于其靶基因的附近，因此 Allen 等认为 miRNA 家族有可能通过基因家族扩增过程中的倒转复制或反向倍增机制（inverted duplication）产生。如图 11 所示，基因家族在扩增过程中由于倒转复制产生头对头或尾对尾的全部或部分基因复制片段，从而为形成 miRNA 编码的发卡结构提供了可能。倒转复制可能直接从基因组上发生也可能通过逆转录后结合类似假基因序列形成。甚至一个基因家族相近的成员间的结合也可以产生这样的创始基因（founder gene）。新形成的位点转录得

到的具有发卡结构的转录本有可能称为 DCL 的靶标而导致 siRNA 的产生，从而使创始基因及其相关的家族成员在转录后水平或染色质水平受到 RNA 干扰机制的调控。部分创始基因在分化过程中因维持发卡结构以及被 DCL 的识别的功能限制，形成一类特异的 siRNA 家族（步骤 2）。而对 DCL1 调控的代谢途径的适应性进化导致了 miRNA 基因的形成（步骤 3）。由于变异的持续积累，部分基因在发卡结构和 DCL1 识别功能限制下，只剩下 miRNA 及其互补的 miRNA*一段与原始的序列相似（步骤 4）。miRNA 座位的复制导致了 miRNA 家族其他成员的产生（步骤 5），并由于变异的积累导致不同成员拥有了各自特异的 miRNA 靶基因。结合 miRNA 靶基因家族的进化使该模型变得更加完整。大多数 miRNA 的靶基因都是一大类基因家族中的亚类。靶基因家族的复制（步骤 6）为调控的多样化提供了基础。在一个新的 siRNA 或 miRNA 编码基因形成后（步骤 2 或 3），家族成员中小 RNA 结合位点的保留（步骤 7）或丢失（步骤 8a）导致了转录后水平调控的分化。同时也许还伴随着转录调控因子的改变（步骤 8b），导致了进一步的调控机制的差异。miRNA 靶基因随后的复制和分化事件（步骤 9）致使不同 miRNA 家族不同成员间拥有了各自专一的靶位点及调控功能。这样，通过 miRNA 和靶基因之间的复制事件，以及结合位点的保留或丢失而形成了一个新的调控网络。

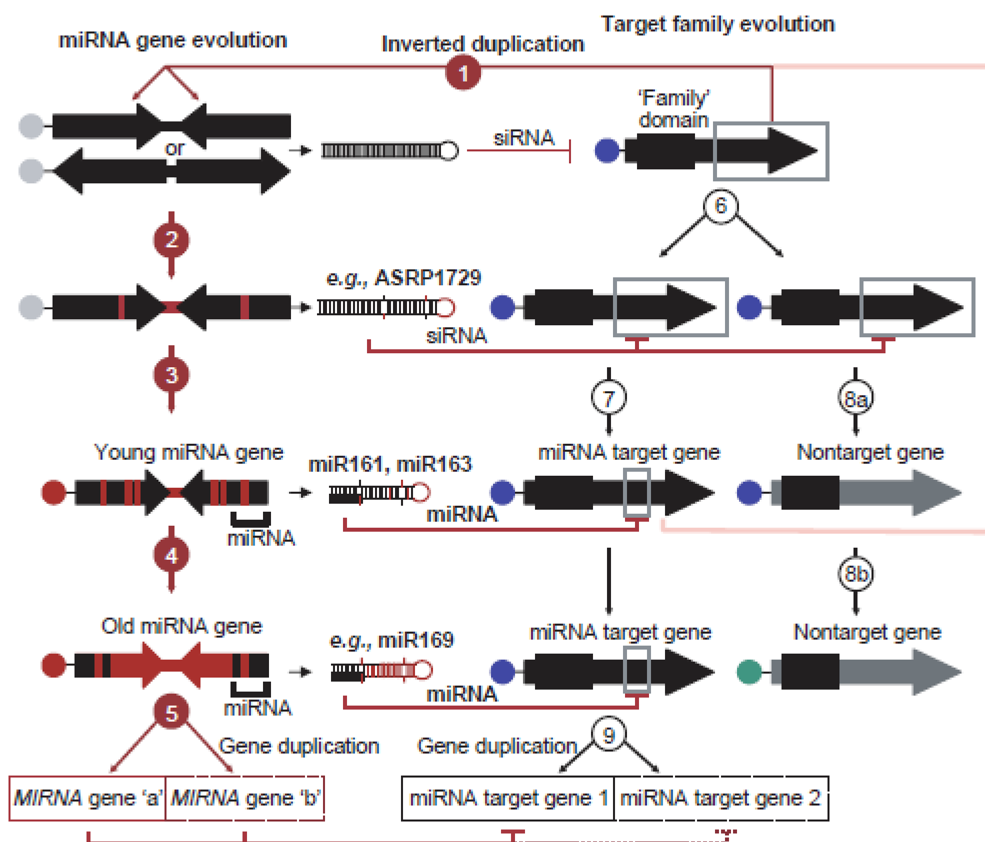


图 11. 植物miRNA反向倍增进化模型(Allen et al. 2004)

然而这样的模型也有很大的局限性。考虑到保守miRNA基因与其靶基因间在结合位点外并没有这种序列相似性的证据存在，对于保守miRNA的解释仍然有待进一步的验证。同样，由于动物miRNA前体序列较短，也不能提供创始基因的信息。一般认为动物miRNA调节机制是通过miRNA和其靶位点间“交互作用获得”事件形成的。跟植物miRNA与靶基因间严格匹配，切割靶基因转录本不同，动物miRNA通过结合到编码基因的3’端干扰其翻译来行使调节作用，并允许其与结合位点间有较多的碱基错配(Bartel et al. 2004)。这一功能模式的不同也表明在动植物miRNA编码基因起源机制上也存在着差异（Li and Mao, 2007)。

对于拟南芥miRNA基因的研究表明，通过上述具有回文结构位点产生的miRNA有几种不同的命运（Fig. 12)：第一，起源于原始基因家族的小RNA保留了调节该基因的能力；第二，小RNA通过遗传漂变获得了特异结合到其他基因或基因家族的能力，很明显，以上两种结果均表明选择作用的存在。第三，也可能是最普遍的命运，随着小RNA产生位点启动子区域、回文结构区域和靶基因结合位点突变的积累而丢失了调节靶基因的能力。因此，植物小RNA的产生机制为研究

特定的调节元件的进化提供了很好的机会（Chapman and Carrington, 2007）。

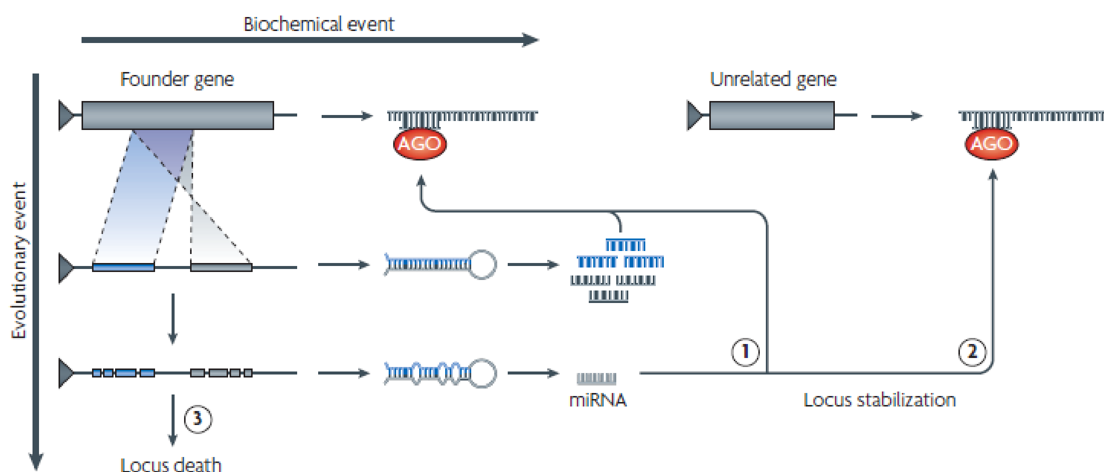


图 12 植物新 miRNA 基因进化模型 (Chapman and Carrington, 2007).

二. 水稻小 RNA 的进化分析

遗传学方面近几年的一个重要的研究进展是在动植物基因组中发现了大量小 RNA 等非蛋白质编码基因，这些小基因（一般 100-200bp）在生理生化等代谢过程中起到重要作用。由此产生一个有待回答的问题：对于水稻等作物中发现的编码小 RNA 的这些基因位点在我们人类进行作物驯化和育种过程中是否同样受到选择（参见第八章）？我们目前在研究作物骨干亲本遗传成因中是否和如何考虑这些基因对骨干亲本形成的影响？目前发现的人工选择（育种）的基因位点主要编码转录调节因子和其他蛋白质编码基因，我们的研究发现非蛋白质编码基因在人工驯化过程中同样受到人工选择效应的影响。我们利用水稻为模式作物，发现小 RNA 之一，miRNA 基因 *MIR156b/c* 基因位点可能受到强烈的自然和人工选择效应的影响，说明人工选择的对象除了转录因子及其下游基因外，还可能针对转录因子调控（上游）基因（Wang et al, 2007）。

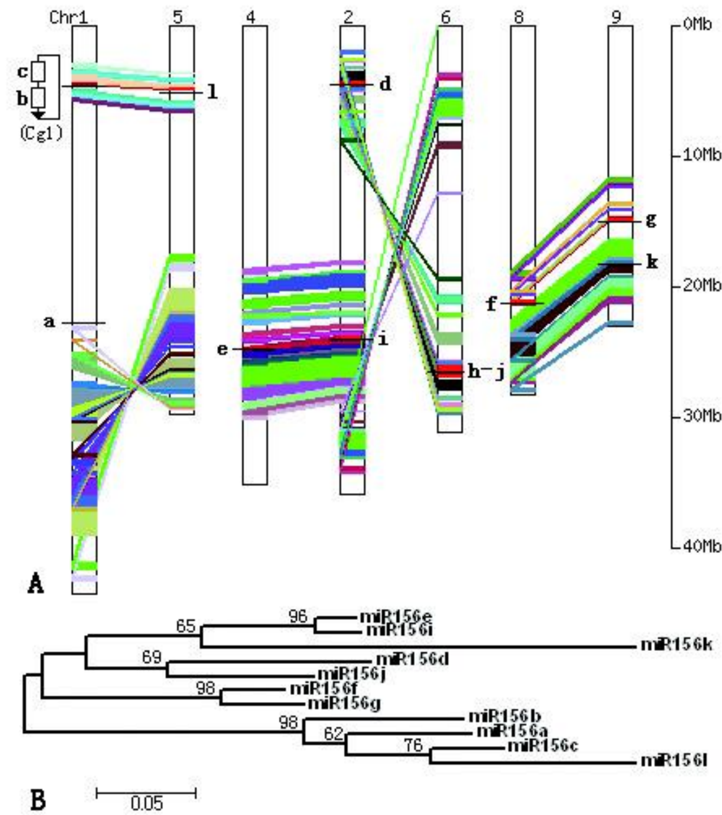


图 13. 水稻 miR156 家族在基因组上的分布和系统进化关系 (Wang et al. 2007)

通过水稻 miRNA 及其靶基因结合位点序列变异的调查和直系同源基因（Paralogs）分析，发现水稻 miRNA 基因在不断地捕获新的结合位点（靶基因），同时也不断丢失对靶基因的调控功能（Guo et al, 2008b）。这种动态的进化过程主要通过 miRNA 序列突变来实现，同时插入和删除也发挥一定作用。图 14 展示了水稻 miR397 靶基因在全基因组前后的突变进化情况，有些靶基因位点由于序列突变而脱离了 miR397 的绑定和调控。

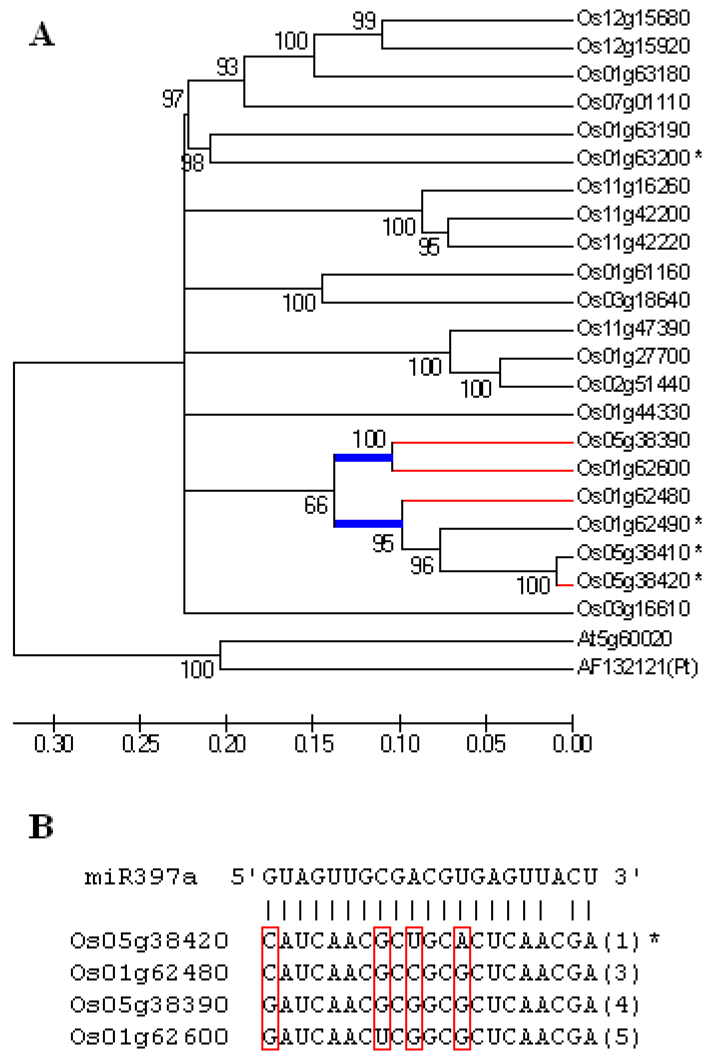


图 14. 水稻 miR397 靶基因进化（A）及其结合位点的序列突变情况（B）(Guo et al. 2008b)

ta-siRNA (trans acting siRNAs)是植物中发现的一类 siRNA 基因(*TAS*)，其在 miR390 等的辅助下，调控生长素相关基因 ARF(auxin response factor)，在植物生长发育过程中发挥重要调控功能。目前已在拟南芥中发现四个亚家族（*TAS1-4*）,其中 *TAS3* 在植物界是保守的。在水稻上，我们通过 Howell 等（2007）和 Chen 等（2007）方面找到了 4 个 *TAS3* 基因（Zhu et al. 2008）。其中部分 21nt 长度读序的 Howell 分布图见图 15。

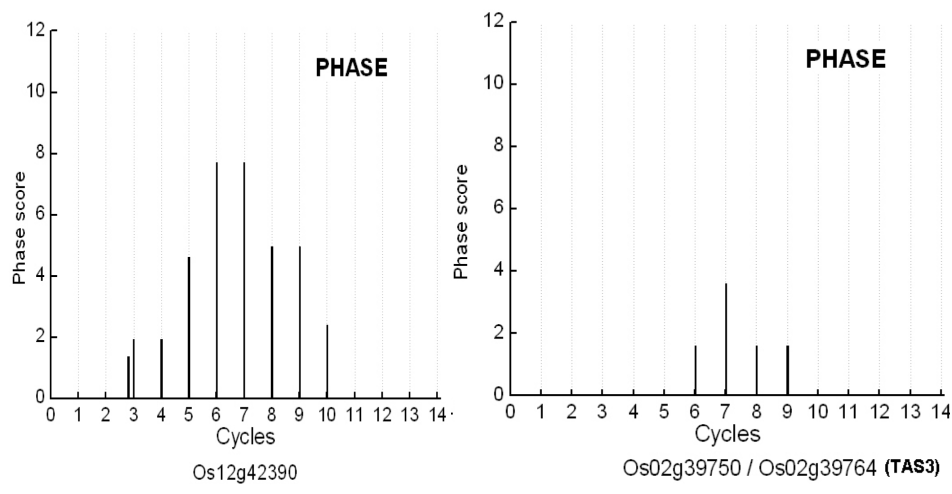


图 15　水稻 *TAS3* 基因 21nt 小 RNA 读序的相位值分布图（小 RNA 数据来自 Zhu et al. 2008）

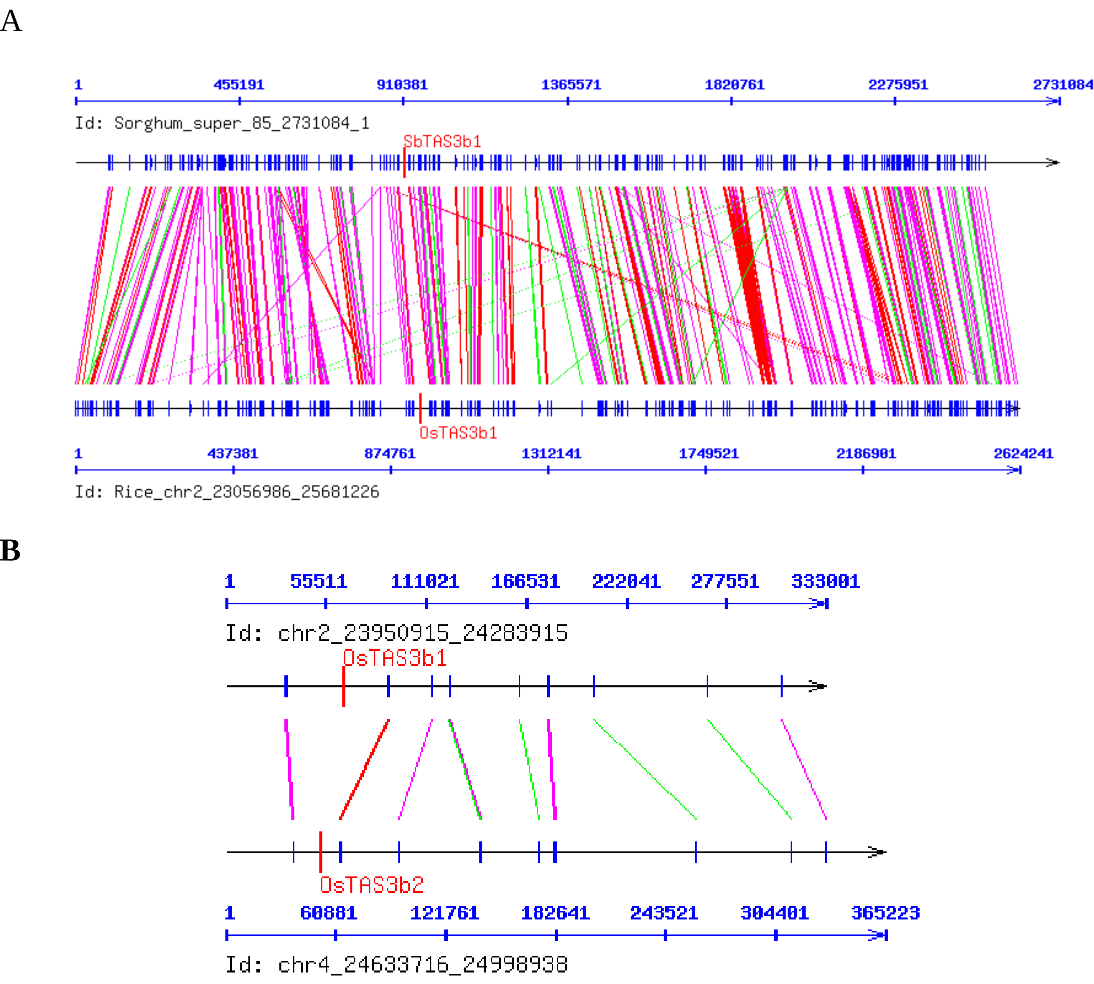


图 16　水稻 *TAS3* 基因倍增及其与高粱同源基因的比较基因组学分析 (Shen et al. 2009)

我们又通过 TAS3 基因的保守序列片段，克隆测序和生物信息学方法发现了 51 个来自禾本科的 *TAS3* 基因（Shen et al, 2009）。通过序列比较等，发现 *TAS3* 基因通过基因组和单基因倍增，在禾本科基因组中至少有 2 个拷贝，多的可达到近 10 个。水稻基因组倍增而来的 TAS3 基因在基因组保持了其共线性关系；同时 TAS3 在不同禾本科基因组上也存在明显的基因组共线性（图 16）。

三. 水稻 miRNA 位点遗传多样性与驯化选择研究

Ehrenreich 和 Purugganan（2008）对拟南芥 miRNA 编码基因及其靶基因的序列变异情况作了大规模调查。通过对 16 个 miRNA 家族 66 个成员及其对应的 52 个靶基因位点的群体数据的分析，表明成熟 miRNA 位点相对于其上下游序列有更高的保守性，并通过中性检验检测到了可能经受选择压力的 miRNA 位点（MiR166f, miR167d, and miR395c）。

为了调查模式作物—水稻中 miRNA 是否经受人工选择即驯化的影响。我们对水稻 miRNA 进行了大规模的群体调查。对 40 个 miRNA 家族的 97 个成员位点进行了重测序，包括了 30 个水稻籼粳亚种的材料。结果表明，与拟南芥的群体调查结果一致，在 miRNA 成熟位点其核苷酸多态性明显低于两端序列，暗示了 miRNA 通过序列互补结合靶基因功能限制的存在。同时，对于保守的 miRNA 家族，其整体的 DNA 多态性相较水稻特异的 miRNA 来说要低一倍，由于保守 miRNA 一般参与基础的代谢网络的调控，因而有可能遭受更强的净化选择而保持序列的保守性(Wang et al. 2010)。

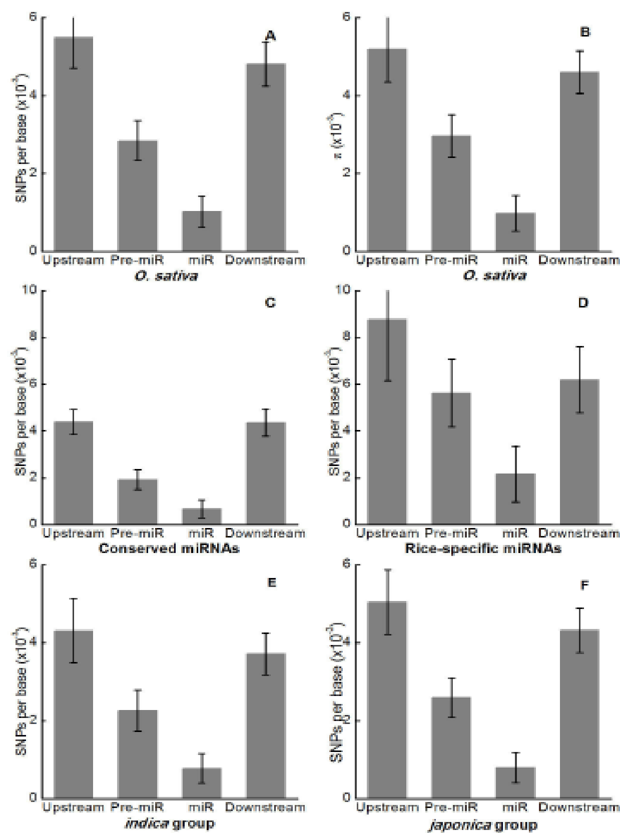


图 16. 水稻 miRNA 位点的序列多态性（Wang et al. 2010）。

另外，我们还对 Tajima's *D* 检验显著的 miRNA 位点进行了进一步的正向选择信号的调查。对相应的 miRNA 位点普通野生稻群体（*O.rufipogon*）进行重测序用于中性检验等分析，结合 Tajima's *D* 检验、HKA 检验的结果，我们找到了几个 miRNA 位点在驯化过程中可能经历了正向选择作用。以 miR390 为例，其调控基因为另一类小 RNA，*TAS3*，中性检验的信号表明，miR390 可能由于选择作用的影响而维持了其特异的调控作用。

第四节 小 RNA 数据库

一. miRBase 数据库

作为目前最权威和完整的 miRNA 数据库 (http://mirdb.org/miRDB/)，截止到目前 (2009 年 11 月)，miRBase 已经收录了一百余个物种中超过 10000 条的 miRNA 记录 (图 17)。其中来自植物体的 miRNA 序列有 1834 条。数据库主要由 3 部分组成：miRBase:Registry，主要是用于提交新的 miRNA 序列；miRBase:Database，用来搜索、比对、下载所有已知 miRNA 相关信息的数据库，包括成熟序列、前体序列、前体二级结构、基因组位置、相关文献等等，并可进行 BLAST 搜索、FTP 下载。miRBase:Targets，存放了所有 miRNA 靶基因的信息。目前已经移至 EBI，并更名为 microCosm。但主要收录了动物 miRNA 的靶基因信息。

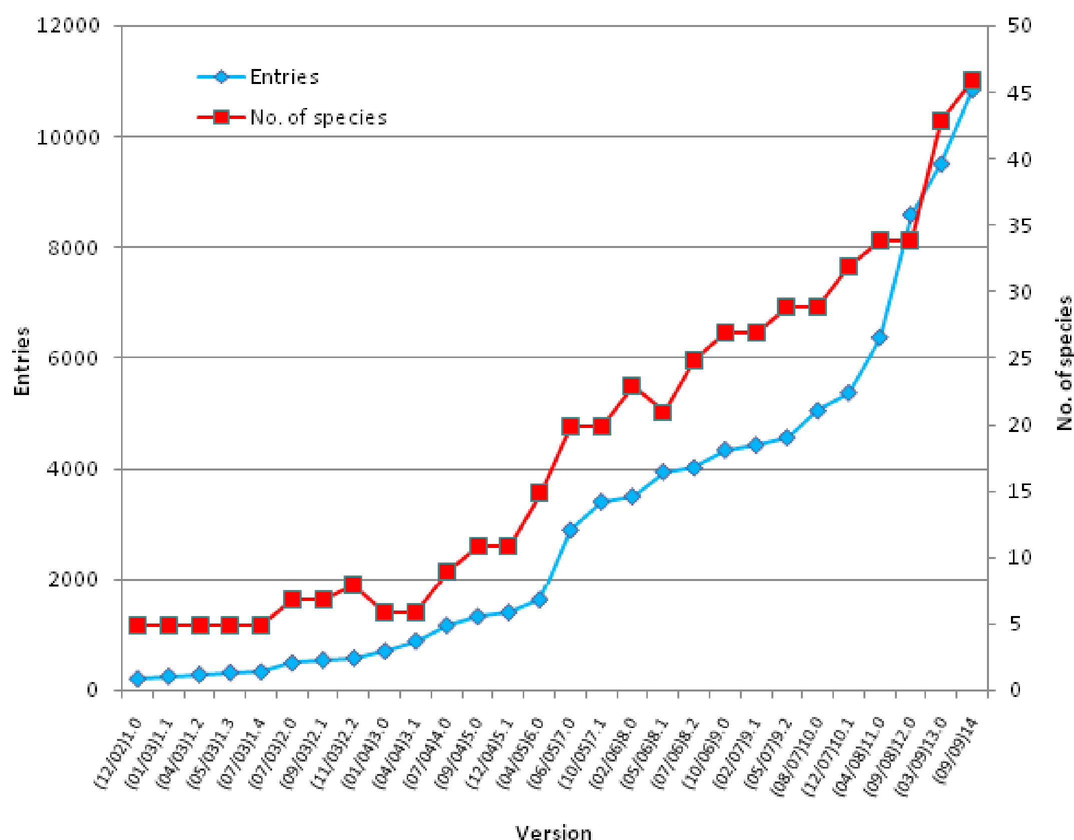


图 17 miRBase 记录和物种数量增长情况

二. siRNA 数据库

由于 siRNA 种类的多样性，为各种类型的 siRNA 建立一个统一的数据库存在很多困难，因此，目前 siRNA 数据的组织没有 miRNA 那样整齐。这里提供两个数据库以供参考，一个是 siRNA Database (http://www.rnainterference.org/)，数据库包括了来自人、大鼠、小鼠的 siRNA 以及 RNAi 等方面的一些资源。另一个是 siRNAdb (http://sirna.sbc.su.se/)，搜集了一千多条经过实验验证的 siRNA 数据和基于计算预测的靶标基因来自 REFSEQ 数据库的 siRNA。

三. CSRDB 和 ASRP

CSRDB (Cereal small RNAs Database, http://sundarlab.ucdavis.edu/smrnas/) 作为专门研究玉米和水稻小 RNA 的数据库，利用 454 测序技术产生了数十万条小 RNA 的数据。可以通过 Genome browser 查看在基因组上的位置信息，并提供了相应的利用 FASTH 软件预测的靶基因数据库 Small RNA target pair (SRTP) dataset。

相应地，ASRP (http://asrp.cgrb.oregonstate.edu/)记录了拟南芥主要生态型和不同组织的小 RNA 数据，包括已知的 miRNA 和 tasiRNA。并提供 BLAST 搜索、Genome Browser 查看、和数据下载。

四. Gene Expression Omnibus (GEO)

Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/gds)作为收录基因表达数据的一个平台，存储了许多原始的表达数据，其中也包括大规模测序的小RNA 数据。大量原始数据的获取，对于从中挖掘小 RNA 研究相关的信息提供了很大的方便。

小结

本章介绍了作为内源性非编码的小 RNA 分子，小 RNA 在最近几年研究的进展。尽管各种新类型的小 RNA 仍在不断地被发现，但依据小 RNA 产生的前体主要分为两类：miRNA 和 siRNA，miRNA 前体可以形成发卡结构，在茎结构处产生成熟的 miRNA，siRNA 主要形成长的双链 RNA，通过各种酶的切割和加工产生成熟序列。植物小 RNA 通过剪切降解靶标 mRNA 分子或在转录后水平干扰翻译来行使调节功能。小 RNA 靶基因一大类是转录因子，miRNA 可以启始 tasiRNA 的剪切。siRNA 类型非常丰富，其中重复序列相关 siRNA 占了很大部分。不同类型小 RNA 的功能研究已经发现了一些结果，但很多疑问还需要深入调查。

生物信息学在计算和数据分析方面的优势决定了其在小 RNA 研究领域所起的重要作用。小 RNA 在序列和结构上存在很多明显的特征，这导致计算方法在不同类型小 RNA 预测，靶位点查找和功能分析方面都取得了卓越的成就。如何利用现有的数据和工具，并开发更加有效更加强大的分析工具是生物信息学人员需要考虑的课题。综合利用不同的数据和方法对提高计算结果的可靠性有重要意义。

可以说作为一个非常重要而且在飞速发展的研究领域，小 RNA 方面的形成机制跟作用机理还有很多的谜团等待着进一步的挖掘。小 RNA 在表达层次表现的功能及复杂性也许正是高等生物进化过程中获得的一个重要的调控机制。小 RNA 序列“身材”上的小巧和通过序列互补调控的机制在生物进化的经济高效方面得到完美体现，并且其中的翻译抑制调节机制是一个可逆的过程，对于生物不断适应变化的生境有着很强的调节机动性。因此，随着研究的深入，不断发现的小 RNA 的新功能和新类型也会将这类 RNA 序列在生物体高效复杂的调控网络中所起的“四两拨千斤”的作用展示得更加令人惊叹！

（王煜，樊龙江）

主要参考文献

[1] Adai A., Johnson C., Mlotshwa S., Archer-Evans S., Manocha V., Vance V..Sundaresan V. (2005) Computational prediction of miRNAs in Arabidopsis thaliana. *Genome Res*, 15(1): 78-91
[2] Allen E., Xie Z., Gustafson A. M., Sung G. H., Spatafora J. W..Carrington J. C. (2004) Evolution of microRNA genes by inverted duplication of target gene sequences in Arabidopsis thaliana. *Nat Genet*, 36(12): 1282-1290
[3] Altuvia Y., Landgraf P., Lithwick G., Elefant N., Pfeffer S., Aravin A., Brownstein M. J., Tuschl T..Margalit H. (2005) Clustering and conservation patterns of human microRNAs. *Nucleic Acids Res*, 33(8): 2697-2706
[4] Ambros V. (2001) microRNAs: tiny regulators with great potential. *Cell*, 107(7): 823-826
[5] Bartel D. P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116(2): 281-297
[6] Baskerville S..Bartel D. P. (2005) Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA*, 11(3): 241-247
[7] Bernstein E., Caudy A. A., Hammond S. M..Hannon G. J. (2001) Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818): 363-366
[8] Bonnet E., Wuyts J., Rouze P..Van de Peer Y. (2004) Detection of 91 potential conserved plant microRNAs in Arabidopsis thaliana and Oryza sativa identifies important target genes. *Proc Natl Acad Sci USA*, 101(31): 11511-11516
[9] Brodersen P..Voinnet O. (2006) The diversity of RNA silencing pathways in plants. *Trends Genet*, 22(5): 268-280
[10] Chan S. W., Zilberman D., Xie Z., Johansen L. K., Carrington J. C..Jacobsen S. E. (2004) RNA silencing genes control de novo DNA methylation. *Science*, 303(5662): 1336
[11] Chapman E. J..Carrington J. C. (2007) Specialization and evolution of endogenous small RNA pathways. *Nat Rev Genet*, 8(11): 884-896
[12] Chen H. M., Li Y. H..Wu S. H. (2007) Bioinformatic prediction and experimental validation of a microRNA-directed tandem trans-acting siRNA cascade in Arabidopsis. *Proc Natl Acad Sci USA*, 104(9): 3318-3323
[13] Dezulian T., Schaefer M., Wiese R., Weigel D..Huson D. H. (2006) CrossLink: visualization and exploration of sequence relationships between (micro) RNAs. *Nucleic Acids Res*, 34(Web Server issue): W400-404
[14] Dunoyer P., Himber C..Voinnet O. (2005) DICER-LIKE 4 is required for RNA interference and produces the 21-nucleotide small interfering RNA component of the plant cell-to-cell silencing signal. *Nat Genet*, 37(12): 1356-1360
[15] Finnegan E. J..Matzke M. A. (2003) The small RNA world. *J Cell Sci*, 116(Pt 23): 4689-4693
[16] Gasciolli V., Mallory A. C., Bartel D. P..Vaucheret H. (2005) Partially redundant functions of Arabidopsis DICER-like enzymes and a role for DCL4 in producing trans-acting siRNAs. *Curr Biol*, 15(16): 1494-1500
[17] Griffiths-Jones S. (2004) The microRNA Registry. *Nucleic Acids Res*, 32(Database issue): D109-111
[18] Griffiths-Jones S. (2006) miRBase: the microRNA sequence database. *Methods Mol Biol*, 342: 129-138
[19] Guo X., Gui Y., Wang Y., Zhu Q. H., Helliwell C..Fan L. (2008) Selection and mutation on microRNA target sequences during rice evolution. *BMC Genomics*, 9:

454

[20] Herr A. J., Jensen M. B., Dalmay T..Baulcombe D. C. (2005) RNA polymerase IV directs silencing of endogenous DNA. *Science*, 308(5718): 118-120

[21] Hofacker I. L. (2003) Vienna RNA secondary structure server. *Nucleic Acids Res*, 31(13): 3429-3431

[22] Howell M. D., Fahlgren N., Chapman E. J., Cumbie J. S., Sullivan C. M., Givan S. A., Kasschau K. D..Carrington J. C. (2007) Genome-wide analysis of the RNA-DEPENDENT RNA POLYMERASE6/DICER-LIKE4 pathway in Arabidopsis reveals dependency on miRNA- and tasiRNA-directed targeting. *Plant Cell*, 19(3): 926-942

[23] Jones-Rhoades M. W..Bartel D. P. (2004) Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell*, 14(6): 787-799

[24] Jones-Rhoades M. W., Bartel D. P..Bartel B. (2006) MicroRNAS and their regulatory roles in plants. *Annu Rev Plant Biol*, 57: 19-53

[25] Kanno T., Huettel B., Mette M. F., Aufsatz W., Jaligot E., Daxinger L., Kreil D. P., Matzke M..Matzke A. J. (2005) Atypical RNA polymerase subunits required for RNA-directed DNA methylation. *Nat Genet*, 37(7): 761-765

[26] Kasschau K. D., Fahlgren N., Chapman E. J., Sullivan C. M., Cumbie J. S., Givan S. A..Carrington J. C. (2007) Genome-wide profiling and analysis of Arabidopsis siRNAs. *PLoS Biol*, 5(3): e57

[27] Kurihara Y..Watanabe Y. (2004) Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc Natl Acad Sci USA*, 101(34): 12753-12758

[28] Lee Y., Kim M., Han J., Yeom K. H., Lee S., Baek S. H..Kim V. N. (2004) MicroRNA genes are transcribed by RNA polymerase II. *EMBO J*, 23(20): 4051-4060

[29] Li A..Mao L. (2007) Evolution of plant microRNA gene families. *Cell Res*, 17(3): 212-218

[30] Lim L. P., Lau N. C., Weinstein E. G., Abdelhakim A., Yekta S., Rhoades M. W., Burge C. B..Bartel D. P. (2003) The microRNAs of Caenorhabditis elegans. *Genes Dev*, 17(8): 991-1008

[31] Llave C., Kasschau K. D., Rector M. A..Carrington J. C. (2002) Endogenous and silencing-associated small RNAs in plants. *Plant Cell*, 14(7): 1605-1619

[32] Lu C., Kulkarni K., Souret F. F., MuthuValliappan R., Tej S. S., Poethig R. S., Henderson I. R., Jacobsen S. E., Wang W., Green P. J..Meyers B. C. (2006) MicroRNAs and other small RNAs enriched in the Arabidopsis RNA-dependent RNA polymerase-2 mutant. *Genome Res*, 16(10): 1276-1288

[33] Nam J. W., Shin K. R., Han J., Lee Y., Kim V. N..Zhang B. T. (2005) Human microRNA prediction through a probabilistic co-learning model of sequence and structure. *Nucleic Acids Res*, 33(11): 3570-3581

[34] Onodera Y., Haag J. R., Ream T., Nunes P. C., Pontes O..Pikaard C. S. (2005) Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell*, 120(5): 613-622

[35] Osato N., Yamada H., Satoh K., Ooka H., Yamamoto M., Suzuki K., Kawai J., Carninci P., Ohtomo Y., Murakami K., Matsubara K., Kikuchi S..Hayashizaki Y. (2003) Antisense transcripts with rice full-length cDNAs. *Genome Biol*, 5(1): R5

[36] Papp I., Mette M. F., Aufsatz W., Daxinger L., Schauer S. E., Ray A., van der Winden J., Matzke M..Matzke A. J. (2003) Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors. *Plant Physiol*, 132(3): 1382-1390

[37] Pontier D., Yahubyan G., Vega D., Bulski A., Saez-Vasquez J., Hakimi M. A.,

Lerbs-Mache S., Colot V..Lagrange T. (2005) Reinforcement of silencing at transposons and highly repeated sequences requires the concerted action of two distinct RNA polymerases IV in Arabidopsis. *Genes Dev*, 19(17): 2030-2040

[38] Rajagopalan R., Vaucheret H., Trejo J..Bartel D. P. (2006) A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev*, 20(24): 3407-3425

[39] Reinhart B. J., Weinstein E. G., Rhoades M. W., Bartel B..Bartel D. P. (2002) MicroRNAs in plants. *Genes Dev*, 16(13): 1616-1626

[40] Rhoades M. W., Reinhart B. J., Lim L. P., Burge C. B., Bartel B..Bartel D. P. (2002) Prediction of plant microRNA targets. *Cell*, 110(4): 513-520

[41] Robins H., Li Y..Padgett R. W. (2005) Incorporating structure to predict microRNA targets. *Proc Natl Acad Sci USA*, 102(11): 4006-4009

[42] Shen D., Wang S., Chen H., Zhu Q. H., Helliwell C..Fan L. J. (2009) Molecular phylogeny of miR390-guided trans-acting siRNA genes (TAS3) in the grass family. *Plant Systematics and Evolution*, 283(1-2): 125-132

[43] Sullivan C. S..Ganem D. (2005) MicroRNAs and viral infection. *Mol Cell*, 20(1): 3-7

[44] Sunkar R., Girke T..Zhu J. K. (2005) Identification and characterization of endogenous small interfering RNAs from rice. *Nucleic Acids Res*, 33(14): 4443-4454

[45] Sunkar R..Jagadeeswaran G. (2008) In silico identification of conserved microRNAs in large number of diverse plant species. *BMC Plant Biol*, 8: 37

[46] Sunkar R..Zhu J. K. (2004) Novel and stress-regulated microRNAs and other small RNAs from Arabidopsis. *Plant Cell*, 16(8): 2001-2019

[47] Tang G., Reinhart B. J., Bartel D. P..Zamore P. D. (2003) A biochemical framework for RNA silencing in plants. *Genes Dev*, 17(1): 49-63

[48] Tran R. K., Zilberman D., de Bustos C., Ditt R. F., Henikoff J. G., Lindroth A. M., Delrow J., Boyle T., Kwong S., Bryson T. D., Jacobsen S. E..Henikoff S. (2005) Chromatin and siRNA pathways cooperate to maintain DNA methylation of small transposable elements in Arabidopsis. *Genome Biol*, 6(11): R90

[49] Vaucheret H. (2006) Post-transcriptional small RNA pathways in plants: mechanisms and regulations. *Genes Dev*, 20(7): 759-771

[50] Vazquez F. (2006) Arabidopsis endogenous small RNAs: highways and byways. *Trends Plant Sci*, 11(9): 460-468

[51] Wang H., Chua N. H..Wang X. J. (2006) Prediction of trans-antisense transcripts in Arabidopsis thaliana. *Genome Biol*, 7(10): R92

[52] Wang J. F., Zhou H., Chen Y. Q., Luo Q. J..Qu L. H. (2004) Identification of 20 microRNAs from Oryza sativa. *Nucleic Acids Res*, 32(5): 1688-1695

[53] Wang S., Zhu Q. H., Guo X., Gui Y., Bao J., Helliwell C..Fan L. (2007) Molecular evolution and selection of a gene encoding two tandem microRNAs in rice. *FEBS Lett*, 581(24): 4789-4793

[54] Wang X., Zhang J., Li F., Gu J., He T., Zhang X..Li Y. (2005) MicroRNA identification based on sequence and structure alignment. *Bioinformatics*, 21(18): 3610-3614

[55] Wang Y., Shen D., Bo S. P., Zheng J., Zhu Q. H., Helliwell C..Fan L. J. (2010) Sequence variation and selection of small RNAs in domesticated rice. *BMC Evol Biol*: Revised

[56] Xie Z., Allen E., Wilken A..Carrington J. C. (2005) DICER-LIKE 4 functions in trans-acting small interfering RNA biogenesis and vegetative phase change in Arabidopsis thaliana. *Proc Natl Acad Sci U S A*, 102(36): 12984-12989

[57] Xie Z., Johansen L. K., Gustafson A. M., Kasschau K. D., Lellis A. D., Zilberman

D., Jacobsen S. E..Carrington J. C. (2004) Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol*, 2(5): E104

[58] Zhang B., Pan X., Cannon C. H., Cobb G. P..Anderson T. A. (2006a) Conservation and divergence of plant microRNA genes. *Plant J*, 46(2): 243-259

[59] Zhang B. H., Pan X. P., Cox S. B., Cobb G. P..Anderson T. A. (2006b) Evidence that miRNAs are different from other RNAs. *Cell Mol Life Sci*, 63(2): 246-254

[60] Zhang B. H., Pan X. P., Wang Q. L., Cobb G. P..Anderson T. A. (2005) Identification and characterization of new plant microRNAs using EST analysis. *Cell Res*, 15(5): 336-360

[61] Zhang Y. (2005) miRU: an automated plant miRNA target prediction server. *Nucleic Acids Res*, 33(Web Server issue): W701-704

[62] Zhu Q. H., Spriggs A., Matthew L., Fan L., Kennedy G., Gubler F..Helliwell C. (2008) A diverse set of microRNAs and microRNA-like small RNAs in developing rice grains. *Genome Res*, 18(9): 1456-1465

[63] Zilberman D., Cao X..Jacobsen S. E. (2003) ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science*, 299(5607): 716-719

[64] Zilberman D., Cao X., Johansen L. K., Xie Z., Carrington J. C..Jacobsen S. E. (2004) Role of Arabidopsis ARGONAUTE4 in RNA-directed DNA methylation triggered by inverted repeats. *Curr Biol*, 14(13): 1214-1220

第八章 遗传多态性及正向选择检测

《物种起源》的发表距今刚好150周年，进化理论的发展也经历了一个漫长而曲折的过程。以哈迪温伯格遗传平衡定律为基础，群体遗传学三巨头：Fisher，Haldane和Wright 建立了群体遗传学的数学基础和理论框架。群体遗传学以数学和统计学的手段研究群体结构的变化，对影响群体结构的因素如环境、遗传变异、遗传漂变、迁移进行了研究。分子群体遗传学在自然选择进化论与Kimura中性进化论的争议声中不断发展。在经典群体遗传学的基础上，以DNA等分子序列为研究对象的分子群体遗传学为种群演化的研究提供了数据来源，并将研究领域扩展到新的层次。伴随着中性理论的挑战和分子群体遗传学的发展，自然选择本身也在不断发展，选择的概念不断细化，正向选择，负向选择，平衡选择等不同进化方式的机制得到深入的研究。尤其是受正向选择位点为揭示一些重要的基因座位的进化历史和遗传动力提供了重要的信息。随着大量分子水平上检测到自然选择作用的证据出现，Ohta对中性理论进行了修改，提出了“近-中性”进化理论（near-neutrality），认为“突变-漂变-选择”三者在分子进化中同时起作用。自然选择在进化中的作用仍然被多数学者所认可。而如今，伴随着群体数据的激增和基因组计划的实施，大规模的正向选择作用的调查表明自然选择在物种形成与进化过程中起着更为重要而广泛的作用。

分子群体遗传学内容丰富，本章将主要从一些基本概念和正向选择的检测方法出发，从群体遗传多态性，变异频率谱线，同义突变和非同义突变比率，单体型和连锁不平衡度，群体结构等群体特征的角度介绍相关的概念，思路，工具和方法。由于分子群体遗传学涉及的知识点很多，发展也很迅速，因此短短的一章内容是无法涵盖这一领域所有的概念或知识，应该再参考《分子进化与系统发育》等群体遗传学领域的专著以及其他相关文献进行更加系统的学习。希望本章的简要介绍能为国内对相关领域感兴趣的学生提供一些参考。

第一节 群体遗传多态性估算

大多数生物的自然群体具有大量的遗传变异。对于一个编码蛋白质的遗传座位，在群体中通常含有两个或多个等位基因（allele）。在一个群体中，存在两个和多个有着相当高频率（通常大于 1%）的等位基因时称为遗传多态性。遗传多态性的产生机制有很多，如核苷酸替代、插入、缺失、转换和等位基因间的重组等。大多数新突变由于遗传漂变或净化选择作用从群体中淘汰掉，只有极少数突变在群体中保留下来。对遗传多态性的产生和维持以及群体水平上的进化机制研究是群体遗传学的主要课题，正如 Kimura 和 Ohta (1971)所指出的，基因的长期进化和遗传多态性仅仅是同一个进化过程中的两个方面。中性学说理论认为，分子水平上的遗传变异在很大程度上是中性的，变异程度主要由突变速率和有效群体大小决定（Kimura and Crow, 1964; Nei, 1987）。因此，通过比较观察到的和预测的遗传变异来验证中性进化这一假说。如果观察和预测值之间的差异显著，就有可能存在某种选择作用。一个群体的遗传多态性通常是指等位基因频率或者核苷酸多态性，两者在选择作用的检验方面都有不同的应用。

一. 影响群体遗传多样性的因素

群体或称种群（population），在进化过程中受各种因素的影响，反映在其遗传结构上就产生了复杂的遗传构成。这些因素包括突变（mutation）、种群历史（demographic history）、遗传漂变（genetic drift）、自然选择（natural selection）、重组（recombination）等等，他们对群体的遗传构成产生不同方面的影响。

一般认为，突变为物种的进化提供了物质基础, 增加了遗传的多样性，是进化的主要动力。

在一个小群体内，因为每个个体的后代存活数量存在差异，而每个个体在同一遗传座位上可能携带不同的等位基因，每代传递到下一代个体的基因频率，会产生较大误差，由这种抽样误差引起群体基因频率的随机变化，叫做遗传漂变。遗传漂变主要受有效种群大小影响, 一般来说, 有效种群越大, 遗传漂变的效应越小。

种群历史（demographic history）主要包括种群扩增、奠基者效应（founder effect）、瓶颈效应（bottleneck）、种群缩减、分割（population subdivision）、种群间的基因交流（gene flow）等等影响种群遗传构成的因素。奠基者效应是指遗传

漂变的一种形式，指由带有亲代群体中部分等位基因的少数个体重新建立新的群体的过程。瓶颈效应可以看做奠基者效应的一种。迁移是指对于一个大种群而言，在每个世代有部分个体迁入从而引起基因频率变化。

自然选择作用于非中性突变上，或者增加有利突变在群体中的频率，或者消除不利的突变，或者以其他的方式对遗传的多样性进行修饰。关于选择的作用在第二节会进行更详细的介绍。

不同的因素互相作用互相影响，形成目前我们观察到的种群的复杂的遗传构成，群体遗传学的一个重要内容就是试图分辨遗传漂变和种群历史跟自然选择尤其是正向选择的效应，从而检测出进化上重要的基因位点。

二. 等位基因频率

一个特定等位基因在某个群体中的相对比例称为等位基因的频率。假设一个座位上有一对等位基因 A1 和 A2，频率分别为 x1 和 x2。在二倍体生物的群体中，该座位共有 3 种可能的基因型，即 A1A1，A1A2，A2A2，频率分别为 X11，X12，X22。一般在随机交配，雌雄配子随机结合的情况下，基因频率和基因型频率的关系为：$X_{11}=x_1^2, X_{12}=2x_1x_2, X_{22}=x_2^2$，这一规律称为哈迪－温伯格（Hardy-Weinberg）定律。固定系数 F 是指对于一个座位上的两个等位基因的基因频率，与 Hardy-Weinberg 定律的偏差。比如：

$$X_{11}=(1-F)x_1^2+Fx_1, X_{12}=2(1-F)x_1x_2, X_{22}=(1-F)x_2^2+Fx_2$$

因此　　$F=(2x_1x_2-X_{12})/(2x_1$

若 $2x_1x_2$ 为随机交配（h）情况下杂合子的预期频率，X_{12} 为群体（h_0）中杂合子的观察频率，则上式可表示为：$F=(h-h_0)/h$

当 h_0 小于 h 时，F 取正值；当 h_0 大于 h 时，F 取负值。Nei 对多等位基因群体的 F 的计算进行了阐述，具体细节参阅 Nei 和 Kimura《分子进化与系统发育》一书。

三. DNA 多态性

对于自然群体的遗传变异研究而言，DNA 序列比蛋白质序列提供了更多信息。首先，对于 DNA 非编码区的遗传变异（内含子，基因间区域）或编码区的同义核苷酸替代只能通过 DNA 序列来研究。DNA 多态性可以用不同的方法来度量，比

较常用的是每个核苷酸座位的分离位点数目和核苷酸多样性（或核苷酸水平杂合度）。

```
    6     21
miR166e TCGAACCAGG CTTCATTCCC C
miR166a TCGGACCAGG CTTCATTCCC C
miR166g TCGGACCAGG CTTCATTCCT C
miR166i TCGGATCAGG CTTCATTCCT C
miR166k TCGGACCAGG CTTCAATCCC T
miR166m TCGGACCAGG CTTCATTCCC T
```

图 1. 群体联配数据

1. 分离位点数目

考虑一个给定的 DNA 区域（座位）并假定从一个群体中抽取 *m* 个拷贝（基因）如果 DNA 区域长度为 *n* (*n* 个碱基)，对于这 *m* 条经过多序列连配的序列，任何有两种或多种碱基的位点被称为分离位点（segregating site)，（图 1）。用 *S* 表示一组数据中的所有分离位点数目。用 *S* 表示所有分离位点数目的总和。没个核苷酸座位（p_s）的分离位点数目为 $p_s = S/n$，*n* 为所研究的序列长度。*S* 和 p_s 很明显取决于样本大小，当 *m* 增大时，它们也增大。在满足无限位点遗传模型条件下，即假设任何一对核苷酸座位之间不发生重组而且新突变总是发生在非分离位点，考虑 p_s 的期望值。进一步假设不存在自然选择而且群体达到突变-漂移平衡，p_s 的期望值可有下式得出：

$$E(p_s) = a_1\theta$$

其中，$a_1 = 1 + 2^{-1} + 3^{-1} + \cdots + (m-1)^{-1}, \theta = 4N\mu$（Watterson， 1975），其中，$N$ 和 μ 分别是有效群体大小和每个位点的突变速率。每个序列的突变速率为 $\nu = n\mu$。很明显，$E(p_s)$ 随 *m* 的增大而增大，p_s 的理论方差为：

$$V(p_s) = E(p_s)/n + a_2\theta^2$$

其中，$a_2 = 1 + 2^{-2} + 3^{-2} + \cdots + (m-1)^{-2}$。因此，$p_s$ 的方差也随 *m* 增大而增大。θ 是一个比 p_s 更基本的遗传变异参数，因为它是突变速率和群体大小的积，并且独立于样本大小。可由下式估算：

$$\hat{\theta} = p_s / a_1$$

$\hat{\theta}$的方差为：　　　　$V(\hat{\theta}) = V(p_s)/c_1^2$

这个等式只有在考虑中性突变且群体大小在进化过程中保持恒定时才是正确的。这里的$\hat{\theta}$有时也写做θ_W。

2. 核苷酸多态性

一个不依赖于样本大小 m 的 DNA 多态性的测度是两个序列间每个位点上核苷酸差异的平均值或是核苷酸多样性。定义为：

$$\pi = \sum_{ij}^{q} x_i x_j d_{ij}$$

其中 q 是等位基因的总数，x_i 是第 i 个等位基因的群体频率，d_{ij} 是第 i 个和第 j 个等位基因间每个座位的核苷酸差异数或替代数。在一个随机交配群体中，π 只是核苷酸水平上的杂合度，可由下式估算：

$$\hat{\pi} = \frac{q}{q-1} \sum_{ij} \hat{x}_i \hat{x}_j d_{ij}$$

或　　$$\hat{\pi} = \sum_{i<j}^{m} d_{ij} / c$$　(*i,j* 指第 *i* 和第 *j* 条序列)

其中 $m, \hat{x}_i, c$ 分别是所研究的 DNA 序列的总条数，样本中第 i 个等位基因的频率和序列比较的总数[$m(m$-1)/2]。

作为θ的两个估计值，θ_W和π的差异反映了群体在核苷酸多态性水平上偏离中性进化且处于突变-漂移平衡的理想模型的程度。另外，衡量 DNA 多态性还有一个很重要的指标，变异的频率谱线（frequency spectrum of variation），是指根据不同变异出现的频率计算的杂合度。

第二节 正向选择的统计检验

一. 自然选择的分类

为了阐明不同类型的自然选择，我们仍然以一对等位基因为例进行解释。假设一个群体开始存在单一的等位基因 A_1，在一个时间点上由于突变引入了另一个等位基因 A_2，那么该群体中总共存在三种基因型 A_1A_1，A_1A_2，A_2A_2，定义每种基因型的适合度分别为 W_{11}，W_{12}，W_{22}，简单来说，基因型的适合度是指携带特定基因型的个体存活的几率。为了更好的理解不同情况下发生的选择情况，我们将绝对适合度转化为相对适合度，三种基因型的相对适合度分别为 1，1 + hs，1 + s (1 +

hs = W_{12}/W_{11}, 1 + s = W_{22}/W_{11})，这样 A_1A_2，A_2A_2 的适合度就转化为 A_1A_1 的适合度来表示。其中 s 和 h 分别指选择系数和杂合效应。s 值的正负以及 h 值的大小决定了选择的类型。如果三种基因型的适合度相等，即 s = 0，那么各种基因型频率维持恒定，在进化上是中性的，否则就有选择发生。当 $0 < h < 1$，会产生定向选择。定向选择会限制群体内的变异，使某种特定的基因频率增加或降低。如果 s < 0，表明等位基因 A_2 是有害的，携带该基因的个体适合度低，从而发生净化选择（或称负向选择，purifying selection or negative selection）使 A_2 在群体中的频率降低。如果 s > 0，表明引入的等位基因 A_2 是有利突变，携带该等位基因的个体更适合生存，那么 A_2 将最终在群体中固定下来，这就是一般意义上的正向选择（positive selection)。另一种针对有利位点的选择为当 s > 0，h > 1 时，杂合基因型有最高的相对适合度，称为超显性选择（也称为杂合子优势，overdominant selection or heterozygote advantage)。超显性选择是平衡选择（balancing selection）的一种。(另一种普遍的观点是认为针对有益位点的选择作用都称为正向选择（Nielsen, 2005），但往往大家关心的都是定向选择范畴的正向选择)。

正向选择通常会造成受选择位点遗传多态性的降低，同时有利变异的积累往往引起选择搭载效应（hitchiking effort）或选择扫荡（selective sweep）(Fig. 图 2)，前者是指对正向选择位点的选择作用会引起相邻连锁位点频率的上升，后者是指受选择位点两侧的序列多态性会因连带效应而保持很低的水平。两种说法其实是一种现象的两种表现，本质是相同的。另外，正向选择往往引起连锁不平衡的增加。连锁不平衡（Linkage disequilibrium, LD）是指不同座位的两个等位基因出现在一条染色体上的频率与随机组合出现的频率不一致的情况。这些特征均是用来检测正向选择的信号。但需要注意的是，随机漂变或种群动态的影响往往也可以引起遗传构成的变化，如何有效的区分不同因素的影响是目前仍需解决的难题和热点。

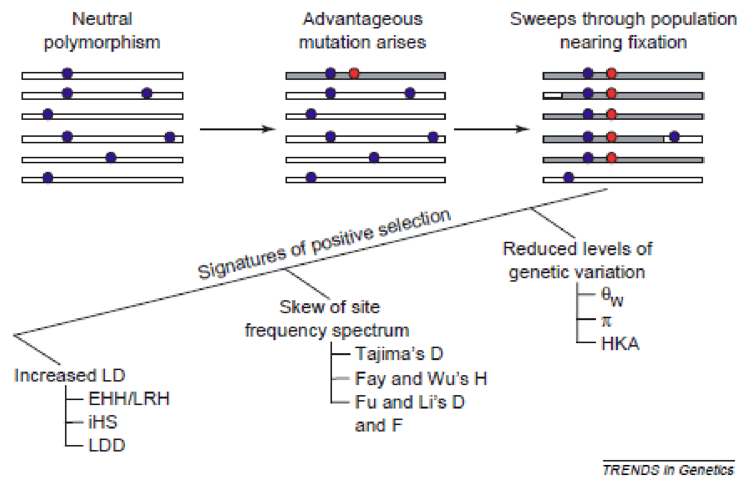


图 2. 受正向选择的等位基因信号与检测方法的关系（Biswas and Akey, 2006）。

二. 中性检验

以中性进化学说作为零假设，通过统计检验的方法检测一个群体的遗传参数是否符合中性进化模型，如果拒绝零假设，表明有其他因素比如选择效应的存在，这类方法统称为中性检验。目前为止，中性检验的方法已经开发了很多，依据利用的数据大体可分为三类：基于种内多态性的检验方法（intraspecific polymorphism）、基于种间分歧度的检验方法（interspecific divergence）和基于种内多态和种间分歧度（intraspecific polymorphism and interspecific divergence）的检验方法。需要注意的是，在具体的分析过程中，一种检验的结果往往不能给出可靠的结果，需要结合多种检验以及具体的生物学背景才能给出比较合理的解释。

1. 基于种内多态性的检验方法

1.1. 基于位点变异的频率谱线

1.1.1. Tajima's *D* 测验

Tajima's D 检验通过比较群体突变率的两个估计值θ_W和π的差异检测正向选择效应。前面提到了群体遗传参数θ的理论值为$\theta = 4N_e\mu$，*Ne* 为有效群体大小，μ为突变频率。然后根据两个估计值θ_W和π的差异构建 Tajima's D 检验：

$$D = \frac{\pi - \theta_W}{\sqrt{V(\pi - \theta_W)}} \qquad \text{（Tajima，1989）}$$

通过蒙特卡罗随机模拟（Monte-Carlo simulation）产生 Tajima's D 检验的分布曲线和临界值，*D* 值的分布并非严格的正态分布，反而与β分布比较接近。实际计算过程中也可以根据实际数据进行模拟进行检验。

在中性进化条件下，θ_W和π的值应该近似相等。因此在标准中性进化模型下，Tajima's *D* 的理论值为零。由于θ_W的计算不考虑分离位点的频率，只跟分离位点的数目有关，所以即使群体中存在大量的低频变异也会对θ_W产生很大影响。由于π计算的是群体中序列差异的平均值，因此π的大小跟变异频率有关。如果实际的 Tajima's *D* 值明显偏离零，表明实际的等位基因频率相对于中性进化模型的期望存在偏倚。如果 Tajima's *D* 值为正，表明存在大量的中等频率的等位基因，这可能是由于群体瓶颈效应，群体结构，或者平衡选择引起的。如果 Tajima's *D* 值为负，表明存在大量的低频等位基因位点，以下几种情况可能会导致 *D* 值为负。首先，当所研究的群体中产生有害突变时，这些突变将受到负向选择的作用在群体中保持较低的频率，低比例的突变有所增加，导致 *D* 值为负。另外，当群体中一条等位基因受到强烈的正向选择作用时，其附近与之紧密连锁的座位的上变异将伴随这条等位基因比例的升高而增加自身在群体中的比例，即选择搭载效应。搭载效应过后，中性突变的积累同样会造成额外的低比例的变异。因此，*D* 值如果为负显著，既可能是负向选择造成的，也可能是正向选择的信号。最后，*D* 值显著并不一定是选择造成的，只是可能存在选择作用的信号。

1.1.2. Fu 和 Li *D* 和 *F* 测验

和 Tajima's *D* 检验相似，Fu 和 Li *D* 和 *F* 检验也是根据等位基因变异频率的偏倚检测群体是否偏离中性进化。所不同的是，后者考虑变异出现的时间因素，即根据在系统进化树上位置确定早期产生的突变与近代产生的突变的分布差异，或根据系统进化树上的位置称为外缘突变（Fig3. d, e, f, g, h）或内部突变（图 3. a, b, c）。

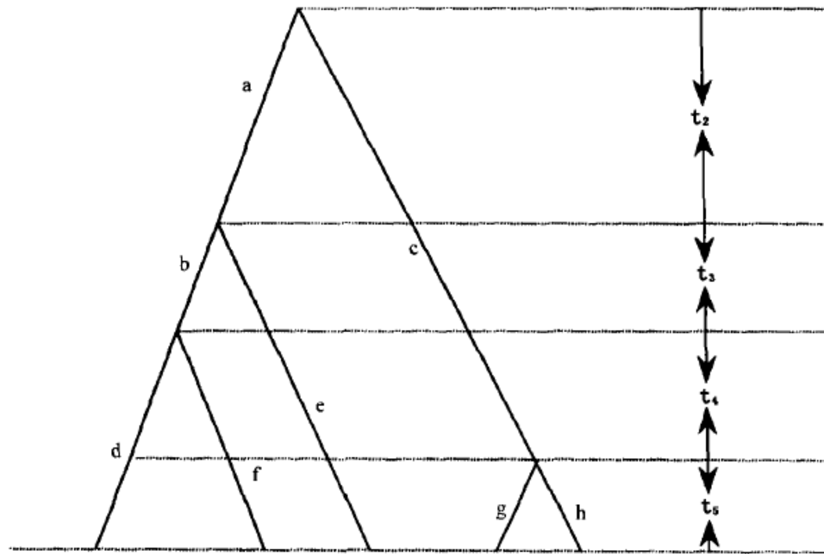


图 3. 一棵 5 条序列构建的系统树（周琦等，2004）。

如果种群受到负选择作用，有害变异频率因选择而降低，或一条有利的等位基因频率受正向选择作用在种群中刚固定不久，都会导致外缘突变相对内部突变的比例大大增加。相反，如果受到平衡选择的影响，则外缘变异相对较少。Fu 和 Li *D* 和 *F* 检验构建了四种统计检验量，不同的检验量之间只是根据不同的方法对θ进行估算，这里只介绍根据外群对θ进行估计的检验。外群（outgroup）是指在进化关系上与所研究种群近缘但又不属于同一类群的分类单元。比如相对于 *O.sativa* 来说，*O.bathii* 可以看做是其外群。利用外群的数据可以构建一颗有根树（rooted tree），计算外缘突变：

$$E(\eta_e) = \theta$$ （Fu and Li, 1993）

内部突变：

$$E(\eta_i) = (a_1 - 1)\theta$$ （Fu and Li, 1993）

其中，$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$，$n$ 为样本数目。

构建统计检验量：

$$G = \frac{\eta_e - \frac{\eta_i}{a_1 - 1}}{\sqrt{V(\eta_e - \frac{\eta_i}{a_1 - 1})}}$$ （Fu and Li, 1993）

类似地，*G* 也近似于β分布。Tajima's *D* 检验和 Fu 和 Li *D* 和 *F* 检验均可以通过 DnaSP 进行计算。

1.1.3. Fay 和 Wu's *H* 测验

如前所述，不同的进化因素往往产生相似或相同的 DNA 多态。比如背景选择效应与搭载效应都会造成种群平均杂合度的降低。此时一些中性检验对如何区分正向选择效应有些力不从心。为了解决这一问题，Fay 和 Wu（2002）提出了一个专门检验搭载效应的中性检验方法：*H* 检验。*H* 检验与 Tajima's D 检验的区别是前者利用通过变异频率估计得到的 θ 的估计值 θ_H 与 π 进行比较。假设样本大小为 *n*，出现过 *i* 次变异的数目为 S_i，那么：

$$\theta_H = \sum_{i=1}^{n-1} \frac{2S_i i^2}{n(n-1)} \qquad \text{(Fay and Wu, 2000)}$$

θ_H 对于高比例的变异比较敏感，当有搭载效应存在时，将产生高比例的变异，这是搭载效应区别背景选择效应的一个显著标志。利用这一特征构建 *H* 检验：

$$H = \frac{\theta_H - \theta_\pi}{\sqrt{V(\theta_H - \theta_\pi)}} \qquad \text{(Fay and Wu, 2000)}$$

当 H 在统计上显著时，表明所研究的种群有可能受搭载效应的影响。*H* 检验可以通过访问http://www.genetics.wustl.edu/jflab/htest.html计算。

1.2. 基于连锁不平衡

在一段DNA序列中，位点与位点之间存在着连锁的关系。不同位点间的连锁构成了“单倍体型”。随着重组的积累，特定的单倍体型会被削弱而逐渐消失。由于重组率与连锁距离有关，所以连锁不平衡范围会逐渐缩短。对于新产生的一个单倍体型，由于重组来不及破坏位点之间的连锁，所以它们之间连锁不平衡的距离往往比较远。在中性条件下，如果某个单倍体型是较新产生的，那么它的频率往往较低，而频率较高的单倍体型，需要经历很长一段时间才可能因为受到随机漂变的影响达到较高的频率。如果群体经历了正向选择，那么与有利位点连锁的周围位点会由于搭载效应频率很快提升，所以包含有利位点的单倍体型一方面有着较高的频率，另一方面由于经历的时间不长，因此也有着较长的LD影响范围。这种特征为检测是否发生了正向选择提供了一个有效的突破点。

1.2.1. LRH (Long range haplotype) 测验

Sabeti 等（2002）提出了 LRH 方法，通过对基因组上的核心单倍体型（Core haplotypes）的研究提出了一种可以进行全基因组扫描的检测正向选择的方法。所谓的核心单倍体型就是指基因中存在的重组率较低的密集区域。计算它们的连锁不平衡度，如果某个核心单倍体型的连锁不平衡程度高于具有其相同频率的一般

单倍体型，那么这个位置很有可能经历了正选择。假如要测量距离核心单倍体型为 x 的区域，其连锁不平衡的衰减通过 EHH（Extended haplotype homozygosity）来计算。EEH 的定义是：两条随机选择的染色体从核心单倍体型到距离为 x 之间的区域存在相同核心单倍体型的概率（图 4）。

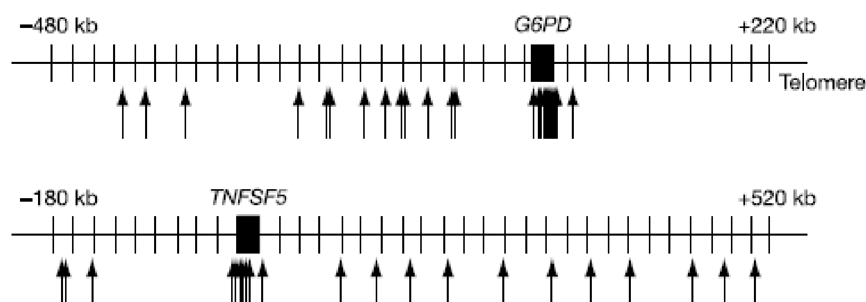


图 4. 以 G6PD 和 TNFSF5 两个位点说明核心单倍体型与周边 SNP（Sabeti, et al. 2002）。

1.2.2. HS (Haplotype similarity)

HS检验是计算单倍体型相似性的检验方法。对于一批DNA样本数据，观察其第一个多态位点，记这个多态位点上频率较低的等位基因为X，然后计算X所关联的染色体的HS值。计算方法是通过一个滑动窗口滑过整段染色体，计算每个窗口中单倍体型的纯合度，然后对所有的窗口取平均值。

$$HS = \frac{\sum_{t=1}^{T}\sum_{i}^{k} f_{it}^{2}}{T}$$

其中T是窗口的总数，k是一个窗口中不同单倍体型的个数，f_{it}是与X相关联的单倍体型的频率（Hanchard et al. 2006）。上述过程是以第一个多态位点为基准进行的，同样可以以第二个、第三个等多态位点为基准进行类似的计算。以某个多态位点为基准计算，如果其相关HS值的水平高于同等频率下的其他多态位点，那么在该多态位点上可能发生了正选择。

1.2.3. iHS (iHH score) 测验

iHS是通过计算同一个SNP上旧的和新的等位基因的iHH比值并取对数得到的：

$$iHS = \ln(\frac{iHH_A}{iHH_D})$$

其中iHH指对EHH的积分（Integrated EHH），A指旧的（Ancestral）等位基因，D指新的（Drived）等位基因（Voight et al. 2006）。iHS的基本原理和LRH很相似。当iHS为较大的正值时，暗示长的单倍体型可能包含旧的等位基因，而iHS为较大的负值时，暗示长的单倍体型可能包含新的等位基因（Biswas et al. 2006）。

1.2.4. LDD (Linkage disequilibrium decay) 测验

LDD 测试指连锁不平衡衰减测试。而 LD 是通过计算 FRC（Fraction of recombinant chromosomes）来体现的。具体方法是，对于一个多态座位，不考虑这个座位上的杂合子，而在纯合子中观察其较少的等位基因和较多的等位基因，考察所有的染色体，将其中与较少等位基因关联的编成一组，而将与较多等位基因关联的编成另一组，然后分别在两组内这个位点周围一个事先预设好的窗口中，计算重组频率与距离的关系，也就是计算不同距离范围内相应的重组率。将这些重组率和相应的距离配对、列表，和标准中性模型的这些值进行方差比较，即计算出 ALnLH（Average log likelihood）（Wang et al. 2006）。在正选择发生时，临近选择位点的 ALnLH 将高出一般的水平（Biswas et al. 2006）。

1.3. 基于群体分化

前面提到，群体的固定系数 F 反映了群体等位基因杂合性水平。固定系数 F 是 F 统计量（F_{ST}）的一个特例，F 统计量，一个比较简单的理解是通过遗传多态性的数据，如 SNP 或微卫星标记，估计亚种群间平均杂合性大小与整个种群平均杂合性大小的差异（$F_{ST} = (H_T - H_S) / H_T$，其中 H_T 代表整个种群平均异质性的大小，H_S 代表亚种群间平均异质性大小）。F 统计量反映了群体结构的变化，它受不同因素的影响，比如突变，遗传漂变，近亲交配，选择作用或 Wahlund 效应（指一个种群中由于亚种群的结构导致的异质性的下降）。在中性进化条件下，F 统计量的大小主要决定于遗传漂变和迁移等因素的影响，如果种群中一个等位基因因为对于特定生境的适合度较高而经历适应性选择，那么其频率的升高会增大种群分化水平，反映在 F 统计量上就是有较高的 F_{ST} 值（$0<=F_{ST}<=1$，F_{ST} 为 1 表示亚种群间存在明显的种群分化）。

2. 基于种内多态和种间分歧度的检测方法

按照中性进化假说的假设，随机遗传漂变是进化的主要动力，因此种内DNA多态性与种间DNA分歧度的进化速率应该一致。如果种内多态性和种间分歧度之间存在显著的偏差，表明种群进化受到了其他因素的影响，暗示了选择作用的存

在。

2.1. McDonald和Kreitman (MK)测验

MK检验的原理是：在无选择作用的中性条件下，所研究基因的种内的同义、非同义突变应与种间同义、非同义突变成正比。反之，则推翻零假设，即基因在不同物种中受到了选择的作用。MK检验思路简洁，计算简单，但在检验中性假说方面却很有说服力。而且该检验与以上提到的检验相比，不需要很多假设限制，重组和种群大小的动态对检验结果没有影响。McDonald和Kreitman(1991)对所研究的DNA序列的位点首先进行分类，以区分种内差异和种间差异。将种内个体间无碱基差异而种间有明显碱基差异的位点，定义为固定位点(fixed site)，作为种间差异的标志。将种内个体间有碱基差异的位点，定义为多态性位点(polymorphic site)，作为种内多态性的标志。分辨出样本的多态位点和固定位点之后，将各位点上的突变再按同义突变位点和错义突变位点加以区分。按照MK检验的原理，在中性条件下：

$$\frac{E(n_f)}{E(s_f)} = \frac{E(n_p)}{E(s_p)}$$

式中n_f代表既是非同义突变位点又是固定位点的位点数，s_f代表既是同义突变位点又是固定位点的位点数，n_p代表既是非同义突变位点又是多态位点的位点数，s_p代表既是同义突变位点又是多态位点的位点数。

当选择作用存在于不同物种中时，上式两边会不相等。此时，可用统计学的*G*-test检验等式两边比例差异的显著性。若显著，也就是说物种间的错义突变数目大于基于种内多态性估计得到的期望值，说明基因在物种间受到了选择作用（周琦等，2004）。根据MK检验的原理可以看出，其应用的范围有限制，即只能对蛋白质编码区进行检测，而且只能利用DNA序列的数据。MK检验可以利用DnaSP软件计算。

2.2. HKA测验

该检验方法基于的原理与MK检验相近，但运用的是统计学的卡方（χ^2）检验。即计算出种间和种内差异的卡平方和，再检验实验结果是否与中性条件下的期望值吻合，所以在统计学上也被称为吻合度检验（goodness of fit test）。

假设K_{1i}代表种1内第i座位DNA序列的分离位点数目，K_{2i}代表种2内第i座位DNA序列的分离位点数目，D_i代表种1和种2间第i座位序列的碱基差异数。将三者

的卡平方和相加得到：

$$\chi^2 = \sum \frac{[K_{1i} - E(K_{1i})]^2}{V(K_{1i})} + \sum \frac{[K_{2i} - E(K_{2i})]^2}{V(K_{2i})} + \sum \frac{[D_i - E(D_i)]^2}{V(D_i)}$$

HKA检验对数据的要求比较高。计算K时需要有两个物种，并需要有两个或两个以上座位的DNA数据。其次该检验要求所研究种群大小保持恒定不变，座位间无连锁。（周琦等，2004）

目前已有很多工作利用HKA检验检测正向选择的信号，而且得到许多可信的结果（Yamasaki et al. 2007; Zhao et al. 2008），表明HKA检验是一种比较有效的方法。基于多位点的HKA检验（Multi-locus HKA test）增加了参照位点的数目，使受检验位点与参照位点的差异更能反映非随机的差异信息，检测结果更加可靠。比如我们在对中国糯玉米群体中的淀粉代谢途径进行驯化信号的调查时（Fan et al. 2009）利用了多位点的HKA检验的方法，选择了6个经证实在玉米群体中进化上是中性的即不受选择作用影响的位点作为参照位点，通过比较待检测位点位点的种内多态和种间差异是否跟参照位点存在显著的统计差异来判断该位点是否存在选择作用。多位点的HKA检验可以通过Hey Lab开发的软件来计算，主要包括SITES和HKA两个软件（http://genfaculty.rutgers.edu/hey/software），首先通过SITES得到每个位点用于HKA检验计算的输入信息，然后利用HKA比较参照位点和待检测位点的差异，通过模拟构建分布给出检验的统计显著值。

3. 基于种间分歧度的检测方法

3.1. Ka/Ks 测验 (*Z* 测验)

自然界中发生的很多非同义突变都是有害突变。在净化选择的作用下这些位点的碱基替换率比较低。假设K_a为非同义突变速率，K_s为同义突变速率。由于同义突变不改变氨基酸序列，因此可假定同义突变为中性突变。在中性条件下，K_a/K_s期望值为1。大部分情况下，DNA序列的K_a/K_s值由于净化选择作用而小于1。但当正向选择作用存在时，某一受正向选择作用的等位基因的K_a/K_s将升高，甚至显著大于1。这时可通过*Z*检验（单侧检验）来判断K_a和K_s之间是否存在显著差异，若K_a显著大于K_s，即为正向选择的标志。计算K_a和K_s的方法有三类：以Nei-Gojobori为代表的进化通路法（Evolutionary Pathway Methods, Nei and Gojobori, 1986），以Li-Wu-Luo为代表的基于Kimura双参数模型的方法（Methods Basedon Kimura’s 2-Parameter Model, Li et al, 1985），和以Yang的密码子替代模型为代表的最大似然

法(Yang and Bielawski, 2000)。其中后两种方法比较常用，Yang的方法可以通过PAML软件包来计算（http://abacus.gene.ucl.ac.uk/software/paml.html）。通过上述方法计算出K_a和K_s后，构建Z检验：

$$Z = \frac{K_a - K_s}{\sqrt{V(K_a - K_s)}}$$（Nei and Kumar, 2002）

如果得到显著的统计检验结果，表明该位点存在选择作用。（周琦等，2004）

4. 溯祖测验（Coalescent simulation, CS）

当代作物在驯化过程中经历了驯化瓶颈（domestication bottleneck）的作用，瓶颈效应导致栽培群体相对于祖先种整体遗传多态性的降低而选择作用往往只针对某个或几个特定的座位。因此可以构建作物的驯化瓶颈效应的模型(图5)，包括祖先群体大小、瓶颈效应的大小（经历瓶颈效应的群体大小与瓶颈效应持续时间的比率）、重组率等参数。在中性进化条件下，该模型的参数可以通过未受到选择作用的位点用模拟的方法进行确定：如果对于几个中性进化的位点，与其有共同祖先的野生种（未经历瓶颈效应）在经过驯化瓶颈效应的模拟后，其群体遗传参数的模拟值与该位点在栽培群体中的观察值在统计检验上一致，表明所选参数符合实际的驯化过程，从而选择该模型用于待检测位点的检验。然后计算在该模型下待检测位点的野生群体经历此强度的驯化瓶颈后，群体遗传参数的模拟值与栽培群体的观察值是否具有统计上的一致性，以分离位点为例，如果栽培群体内的观察到的分离为点显著低于通过模拟得到的分离位点数，或位于通过模拟得到的分离位点分布曲线的置信区间外，表明该位点除了经历驯化瓶颈效应外，还经受了其他作用的影响，暗示了该位点可能受到了选择作用的影响。

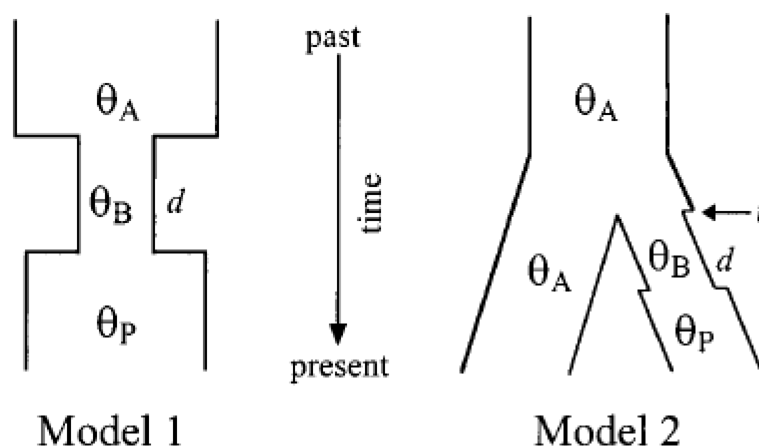


图 5. Eyre-Walker et al. 在调查玉米驯化过程中构建的驯化瓶颈模型(Eyre-Walker et al. 1998)。

5. 复合检验

在各种检验正选择的统计量中，都有各自的优势或劣势，例如Fay 和Wu *H*测验是基于高频突变丰度的检验，它能够比较特异性地检验正向选择，而受群体历史和背景选择的干扰较少，但是它只能检测到刚固定不久的正向选择，因为高频突变将随着时间流逝很快因为随机漂变作用而被固定。Tajima’s *D*在检验正向选择的同时容易受到群体历史和背景选择的干扰，但是Tajima’s *D*所检验的低频突变丰度的信号能够在选择发生位点被固定后持续较长一段时间。一个比较容易想到的方法就是同时利用两种或多种检验方法，使它们的优缺点得以互补，从而能够较特异性地检验正选择。

Zeng等（2006）提出的DH检验就是直接结合了Tajima’s *D*和一个修正后的Fay and Wu’s *H*检验，其检验正向选择的特异性能力相对较高，而对种群历史等其他因素的敏感度很低。后来考虑到Ewens-Watterson的*EW*检验对重组率的变化不敏感，Zeng等（2007）又提出了Fay and Wu’s *H*和*EW*结合的*HEW*检验以及*DH*和*EW*结合的*DHEW*检验，它们相对于*H*检验或*DH*检验对重组率更不敏感。（林栲等，2009）

三. 全基因组扫描及假阳性

基于全基因组重测序的基因组群体遗传学（Hedges, 2000 and Black IV et al. 2001）是大规模检测正向选择位点的一个发展方向，对全基因组的重测序解决了目前研究的几个关键问题：1.单位点的正向选择检测研究往往要求对所研究的位点有一个预先判断，即从基因功能等信息上判断该位点是否有可能受到正向选择，这就导致以往的研究对象总是集中在特定的基因或特定功能以及代谢通路上的基因，而且那些远离蛋白编码区起调节作用的位点往往不能得到很好的研究。比如在Wang等（2007）的课题组对水稻miRNA的群体遗传学研究中发现，作为位于调控网络上游在基因的转录及转录后调控过程中起作用的编码miRNA的基因也检测到了正向选择的信号。而全基因组的扫描可以解决掉单位点研究存在的偏倚，并对非编码区也可以进行调查；2.判断一个位点是否受到正向选择的影响往往需要排除掉种群历史因素的影响，由于种群历史将会影响整个基因组的DNA变异模式，而正向选择只是特异性地作用于某个座位（Black IV et al. 2001），因此对于多座位检验（Multi-locus test）或全基因组扫描的方法，那些与普通状态的座位存在明显差异的位点有可能经历了正向选择。3.由于全基因组扫描调查了大量的位点，因此

可以对某个群体遗传参数比如DNA多态性、变异频谱、连锁不平衡等构建分布，即经验分布（Empirical distribution），将位于分布尾端的异常值（Outlier）看做受到正向选择的候选位点(图6)。

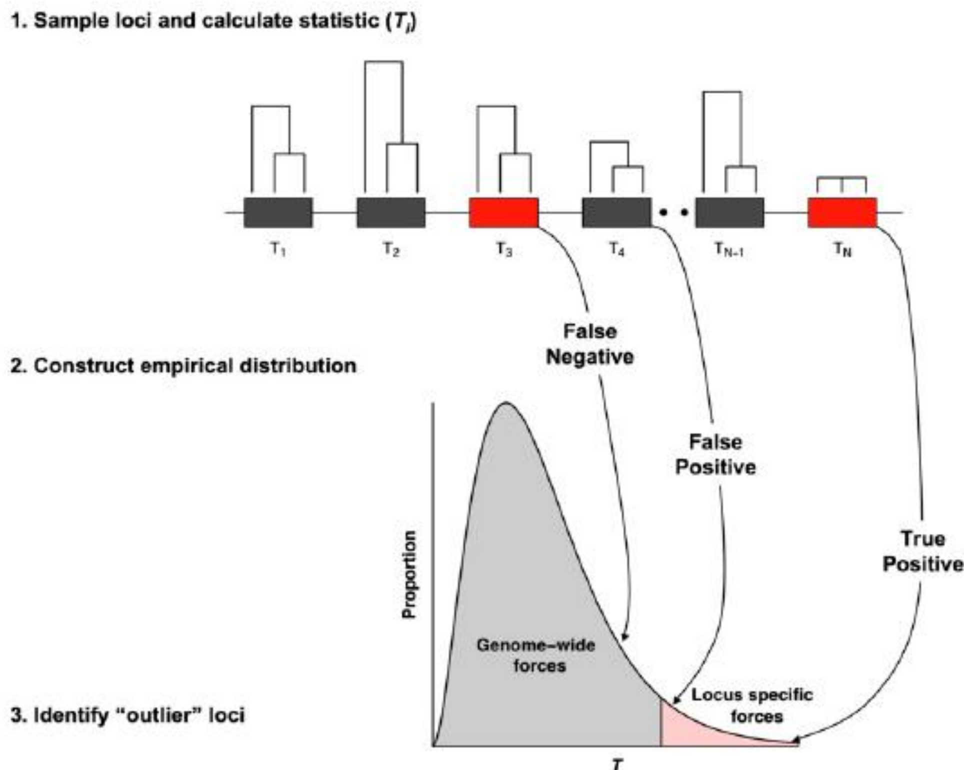


图 6. 对全基因组进行正向选择信号扫描的一般思路（Akey, 2009）。

但是这种方法基于几个假设，而这些假设本身仍存在一些问题，首先，目前还没有证据表明，位于分布尾端的座位一定是经历了正选择的座位，也没有证据表明经历了正选择的座位一定会位于分布的尾端。在经过全基因组扫描过后得到的座位还需要继续用其他的方法进行进一步验证。第二步检验的结果可能会存在确认偏倚(Ascertainment bias, Thornton and Jensen, 2006)如果不考虑这个问题结果将可能有较高的假阳性。另外，不论一个群体是否经历了正向选择，这样计算出来的统计量分布都会存在一个尾部，如果总是把这样的尾部当作正向选择的信号，那么将可能产生假阳性。而且这个尾端占整个分布的比例也很难确定，因为对于非平衡的群体，其经过了选择连带后的座位，在统计量分布尾端以何种程度出现是未知的。为了解决这些问题，一些修正的方案也逐渐被提出来。Thornton等(2006)提出了修正确认偏倚问题和非平衡群体问题的方法，指出如果不考虑确认偏倚的问题，那么第二步检验用到的似然率检验方法可能会受到第一次扫描选择出来座

位的影响很大。同时他认为在选择统计量的时候，采用多样性和群体分化程度的统计量相对于选择变异谱线更为有效，而且较大的基因组区域（2.5kb）对于确定正向选择位点具有更大的功效。

在全基因组扫描中，因为选择模式的不同和受种群历史的干扰，也会产生假阳性问题。例如，如果选择作用在隐性等位基因而不是共显性等位基因上，或者选择作用在一个新的突变而不是一个已经存在的分离位点上，或者选择的同时经历了瓶颈效应，这些因素都可能会增加假阳性（Teshima et al. 2006）。

此外，对于基于高频突变的Fay and Wu’s *H*测验以及基于连锁不平衡的统计量，都会在正选择完成后不久很快丧失其特征，因为高频突变会很快被固定，而连锁不平衡会很快被重组打断(Przeworski, 2002)。如果采用基于高频突变变化或基于连锁不平衡的统计量来进行全基因组的扫描，可能会遗漏一些可能经历过正选择的座位。

另一个策略是用基因组水平的多态位点估计种群历史变化参数，再将估计的参数作为原假设（Null hypothesis），通过似然率测试（Likelihood ratio test）来检测正向选择(Li and Stephan, 2006)（林栲等，2009），即溯祖方法的应用。

针对目前各种检测正向选择作用的方法，综合利用不同的方法对提高检测的有效性和降低假阳性是必须的。因此，在检测正向选择的时候，可能需要同时考虑很多因素，以下任何一个因素的改变都可能使得预测的结果发生偏差:

1. 种群历史、背景选择和平衡选择：在这些检验方法中，一个比较普遍的问题是如何区分自然选择和种群历史的作用。一个思路是，自然选择往往作用于某些特定的座位，而种群历史影响整个 DNA 序列。如果某个区域具有偏离了中性进化的特征，那么要判断这是因为经历了自然选择还是因为种群历史的变化，可以通过将待检测的区域与整个 DNA 序列的相关特征(例如多态性)进行比较，观察其是否和普遍的水平一致，如果不一致的话，那么其可能是经历了自然选择，全基因组扫描就是基于这样一个思路。另一个思路就是尽量考虑自然选择所形成的特异性特征，而这些特征往往是种群历史因素不能形成的，似然率检验就是这样实现的。

背景选择也会对正选择有一定的干扰作用，它们都能产生大量的低频突变，但是正选择产生相对较多的高频突变，这个特征是背景选择没有的。

平衡选择在刚开始的时候和正选择很类似，因为二者在刚开始作用于一个新

的突变的时候都会使这种碱基的频率增加，所不同的是平衡选择最终使新的突变和原有碱基的频率达到一个平衡，而正选择最终会让新的有利碱基替换掉原有碱基。

2. 选择发生的时间：选择是正在进行中的、刚刚完成的还是已经完成很久了，其产生的结果可能是不同的。选择发生的时间不同，所造成的 DNA 变异模式也会有所不同。正选择留下的痕迹可能随着时间的推移而减弱，并且不同的痕迹所能持续的时间也不同。搭载效应产生的高频突变和连锁不平衡就会很快被随机漂变和重组效应消除，而功能(非同义)突变与非功能(同义)突变之间的比例改变就能够持续相当长的时间。在时间上，对于不同的物种来说，群体的真实历史也是需要考虑的一个因素。例如人类从非洲走出的时间大约在 50000 到 75000 年前，所以在人类的 DNA 数据中，如果考察从亚群体间差别程度(例如 F_{st})来检验人类走出非洲后是否经历了某些自然选择，需要考虑到这个时间的因素。

3. 选择发生的位置：突变率和重组率是多少、多座位检验时各个座位的突变率和重组率是否一致，都可能会影响到检验的准确性。不同的突变率和重组率可能会使得检验统计量的临界值发生变化，而多座位检验时，如果各个座位的突变率和重组率不一致，那么可能产生假阳性的结果。另外，在一个位点发生了选择后，其所影响的周围中性位点的距离也是需要考虑的，距离越远，连锁程度越低，影响越弱。在空间上，较早期的检验，例如Tajima *D*检验，都只是考虑了一段DNA序列区域内是否经历了自然选择，而没有考虑自然选择直接作用的位点，而较近期的似然率检验和单倍体连锁不平衡检验，都考虑到了选择直接作用位点及其周围的特异性变异模式(多态性低谷或连锁不平衡衰减)，所以相对灵敏度更高，特异性更强。例如，HS检验的灵敏度与LRH差不多，但是远远高于Tajima *D*等检验方法（Hanchard et al. 2006）。

4. 选择作用的对象：选择作用于新的突变还是已经存在的分离位点，产生的结果可能是不同的。选择作用于一个新的突变上时相对比较容易检验，而如果选择作用于已经存在的分离位点上，那么有的统计量，例如 *H* 和基于连锁不平衡的信号就不是特别强烈了（Przeworski et al. 2002）。

5. 有利等位基因的类型：有利等位基因可能是隐性的，也可能是是共显性的。隐性的有利等位基因产生后，比起共显性的有利等位基因的频率上升的速度相对较慢，选择的力量相对较弱，需要较长一段时间完成选择（Teshima et al. 2006）。

所以在考虑这个因素的时候实际上也是在考虑时间上的因素。

6. 选择的强度：选择的强度可能由多种因素决定，上面提到的等位基因是隐性或共显性也属于影响选择强度的一种因素。选择强度最根本还是取决于自然环境对表型的影响能力。选择强度越强，选择所经历的时间越短。

7. 单个选择还是多次选择：在同一个座位只发生了一次选择事件还是多次，也会造成结果的不同。如果同一个座位经历了多次选择事件，那么在全基因组扫描中， *H* 检验或基于连锁不平衡的统计量就不会表现出那么明显的特异性了（Przeworski et al. 2002）。（林栲等，2009）

四. 研究案例

下面根据我们近期对中国糯玉米淀粉代谢途径的研究说明以上选择检测方法的应用。我们利用糯玉米群体材料和多种检测正向选择的中性检验方法（Tajima's *D*; Fu and Li's D* and F*; KHA; CS; *Fst*），对淀粉代谢途径中的若干关键基因进行了分析（Fan et al. 2008; Fan et al. 2009）。Whitt 等（2002）以普通玉米为遗传材料对淀粉合成代谢途径的六个关键基因 *sh1*、*sh2*、*bt2*、*ae1*、*su1*、*wx1* 做了研究。研究结果表明 *bt2*、*ae1*、*su1* 存在明显的受到正向选择的证据，然而作为控制直链淀粉合成的关键基因，*wx1* 并没有检测到受到正向选择的信号。而糯玉米与普通玉米的一个重要区别便是其表观直链淀粉含量较低（$<5\%$）。基于初步对 30 个糯玉米材料的 *Waxy* 基因位点研究（Fan et al. 2008），我们发现在中国糯玉米群体中，相对于中性位点 *Adh1*，*Waxy* 基因的遗传多态性下降了三到四倍，而且两个中性检验 Tajima's *D* 和 Fu and Li's *D** and *F** 也都检测到了显著的定向选择的信号。在普通玉米中遗传多态性并没有显著的下降，中性检验的结果也不显著。Olsen 等（2006）在糯稻中发现 *Waxy* 基因位点存在明显的选择连带效应，表明强烈的选择作用对该座位有显著的影响。我们同样调查了糯玉米群体中 *Waxy* 基因位点是否存在选择连带效应。分别对 *Waxy* 基因位点上下游的基因进行了群体遗传学的调查，结果表明在 *Waxy* 基因位点上游基因的遗传多态性也维持了很低的水平，表明选择连带效应的影响范围至少延续到 *Waxy* 基因位点上游 50Kb 的位置。

为了进一步了解糯玉米群体中 *Waxy* 基因位点受到正向选择后整个淀粉代谢途径的进化情况，我们对 Whitt 等研究的六个关键基因在糯玉米中进行了群体调查（Fan et al. 2009）。核苷酸多态性的结果表明，相比较其他位点，*Waxy* 基因位点多

态性与普通玉米相比有显著的下降（24.9 倍）。Tajima's *D*、*HKA*、*CS* 检验的结果一致表明 *Waxy* 基因位点受到了强烈的正向选择。值得说明的是 *CS* 检验。由于 Tajima's *D* 和 *HKA* 检验均是以中性进化模型为前体假设，从而不能排除种群历史对检测结果造成的影响。根据以往普通玉米的驯化瓶颈效应的研究，普通玉米驯化瓶颈效应的强度（驯化瓶颈期间群体大小与驯化瓶颈持续时间的比值）约为 2.0~4.5（Zhao et al. 2008; Tenaillon et al. 2004）。而中国糯玉米从明朝由北美洲引入我国，相对于普通玉米存在更为强烈的驯化瓶颈效应，因此我们利用糯玉米的一系列群体遗传学参数进行了 Coalescent simulation 检验，从而排除了种群历史对检测选择信号的影响。特定群体内的选择效应往往产生明显的种群分化的信号（Yu et al. 2008）。*Fst* 检验便是利用种群内不同亚群的分化情况来检测选择的信号。我们以四个中性基因做为参照，对中性检验检测到信号的几个基因（*ae1*、*bt2*、*wx*）通过 bootstrap 获得了其 F 统计量的频率分布。结果表明 *Waxy* 基因位点 F 统计量明显偏向于 1（单边 Kolmogorov-Smirnov 检验，$P < 2.2 \times 10^{-16}$）。表明 *Waxy* 基因位点上糯玉米群体和普通玉米群体显著的分化信号，而其他位点并没有检测到明显的分化证据。

以上研究结果表明，作物突变等原因导致的基因型变异而引起新的表型，会由于后续的驯化或遗传改良的影响而在群体中积累，表现为强烈的正向选择的信号，并可能使一个代谢途径的选择情况发生改变。从分子水平上讲，一个代谢途径中受选择靶标的改变会导致一个重要的农艺性状的迅速积累从而获得携带我们目标性状的品种。

小结

分子群体遗传学是研究种群结构特征的学科，复杂的进化历程在物种的基因组上留下了不同的印迹。受正向选择作用的座位由于往往跟对特定的生境的适应、新功能的获得、物种的进化息息相关，因此如何利用不同方法检测受正向选择作用的位点是群体遗传学一个非常重要的研究内容。中性进化理论坚实的数理统计基础为这一方向提供了强大的工具，并发展出了各种具有严谨的理论假设的检验方法，统称为中性检验方法。不同方法是根据选择对群体遗传参数产生的不同方面的影响而构建的，比如遗传多态性的降低、遗传变异谱线的变化、选择连带效应、连锁不平衡的增加等，这些方法本身基于一定的假设，如果实际情况与这些

假设相背离的时候，这些方法可能会产生假阳性的结果。伴随着测序技术的发展，基于全基因组扫描的群体基因组学成为目前研究的一个趋势，由于基于单位点的正向选择位点检测存在明显的缺陷，因此基于全基因组的受正向选择位点扫描无论从功效上还是检测的有效性都有明显的改进。当然，全基因组扫描仍然存在一些问题，这些需要进一步的研究来提高检测的效率并降低假阳性的比例。

在未来的发展方向上，由于越来越多 DNA 数据的积累，人们将根据具体的数据应用相应的检验方法来推测各种历史事件的发生。另一方面，由于考虑的因素越来越多，各种模型将会越来越复杂，各种检验将会在灵敏度和特异性上不断地努力，这个领域内的理论发展也将越来越具有挑战性。

（王煜，樊龙江）

主要参考文献

1. Akey J. M. (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19: 711-22
2. Biswas S., Akey J. M. (2006) Genomic insights into positive selection. Trends Genet 22: 437-46
3. Black W. C. IV, Baer C. F., Antolin M. F., DuTeau N. M. (2001) Population genomics: genome-wide sampling of insect populations. Annu Rev Entomol 46: 441-69
4. Eyre-Walker A., Gaut R. L., Hilton H., Feldman D. L., Gaut B. S. (1998) Investigation of the bottleneck leading to the domestication of maize. Proc Natl Acad Sci USA 95: 4441-6
5. Fan L. J., Quan L. Y., Leng X. D., Guo X. Y., Hu W. M., Ruan S. L., Ma H. S., Zeng M. Q. (2008) Molecular evidence for post-domestication selection in the Waxy gene of Chinese waxy maize. Molecular Breeding 22: 329-338
6. Fan L.J., Bao J.D., Wang Y., Yao J.Q., Gui Y.J., Hu W.M., Zhu J.Q., Zeng M.Q., Li Y., Xu Y.B. (2009) Post-domestication selection in the maize starch pathway. PLoS One 4: e7612
7. Fay J. C., Wu C. I. (2000) Hitchhiking under positive Darwinian selection. Genetics 155: 1405-13
8. Fu Y. X., Li W. H. (1993) Statistical tests of neutrality of mutations. Genetics 133: 693-709
9. Hanchard N. A., Rockett K. A., Spencer C., Coop G., Pinder M., Jallow M., Kimber M., McVean G., Mott R., Kwiatkowski D. P. (2006) Screening for recently selected alleles by analysis of human haplotype similarity. Am J Hum Genet 78: 153-9
10. Hedges S. B. (2000) Human evolution. A start for population genomics. Nature 408: 652-3
11. Hudson R. R., Kreitman M., Aguade M. (1987) A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153-9
12. Kimura M., Crow J. F. (1964) The Number of Alleles That Can Be Maintained in a Finite Population. Genetics 49: 725-38
13. Li H., Stephan W. (2006) Inferring the demographic history and rate of adaptive substitution in Drosophila. PLoS Genet 2: e166
14. Li W. H., Wu C. I. (1985) A new method for estimating synonymous an d nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2: 150-74
15. McDonald J. H., Kreitman M. (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351: 652-4
16. Nei M. (1987) Molecular evolutionary genetics. Columbia University Press, New York,150-71
17. Nei M., Gojobori T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418-26
18. Nei M., Kimura S. (2002) Molecular Evolution and Phylogenetics. Oxford University Press, London, 258-64
19. Nielsen R. (2005) Molecular signatures of natural selection. Annu Rev Genet 39: 197-218
20. Olsen K. M., Caicedo A. L., Polato N., McClung A., McCouch S., Purugganan M. D. (2006) Selection under domestication: evidence for a sweep in the rice waxy genomic region. Genetics 173: 975-83
21. Przeworski M. (2002) The signature of positive selection at randomly chosen loci. Genetics 160: 1179-89

22. Sabeti P. C., Reich D. E., Higgins J. M., Levine H. Z., Richter D. J., Schaffner S. F., Gabriel S. B., Platko J. V., Patterson N. J., McDonald G. J., Ackerman H. C., Campbell S. J., Altshuler D., Cooper R., Kwiatkowski D., Ward R., Lander E. S. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-7
23. Tajima F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-95
24. Tenaillon M. I., U'Ren J., Tenaillon O., Gaut B. S. (2004) Selection versus demography: a multilocus investigation of the domestication process in maize. Mol Biol Evol 21: 1214-25
25. Teshima K. M., Coop G., Przeworski M. (2006) How reliable are empirical genomic scans for selective sweeps? Genome Res 16: 702-12
26. Thornton K. R., Jensen J. D. (2007) Controlling the false-positive rate in multilocus genome scans for selection. Genetics 175: 737-50
27. Voight B. F., Kudaravalli S., Wen X., Pritchard J. K. (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72
28. Wang E. T., Kodama G., Baldi P., Moyzis R. K. (2006) Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc Natl Acad Sci U S A 103: 135-40
29. Wang S., Zhu Q. H., Guo X. Y., Gui Y. J., Bao J. D., Helliwell C., Fan L. J. (2007) Molecular evolution and selection of a gene encoding two tandem microRNAs in rice. FEBS Lett 581: 4789-93
30. Watterson G. A. (1975) On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7: 256-76
31. Whitt S. R., Wilson L. M., Tenaillon M. I., Gaut B. S., Buckler E. S. th (2002) Genetic diversity and selection in the maize starch pathway. Proc Natl Acad Sci U S A 99: 12959-62
32. Yamasaki M., Tenaillon M. I., Bi I. V., Schroeder S. G., Sanchez-Villeda H., Doebley J. F., Gaut B. S., McMullen M. D. (2005) A large-scale screen for artificial selection in maize identifies candidate agronomic loci for domestication and crop improvement. Plant Cell 17: 2859-72
33. Yang Z., Bielawski J. P. (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496-503
34. Yu Y., Tang T., Qian Q., Wang Y., Yan M., Zeng D., Han B., Wu C. I., Shi S., Li J. (2008) Independent losses of function in a polyphenol oxidase in rice: differentiation in grain discoloration between subspecies and the role of positive selection under domestication. Plant Cell 20: 2946-59
35. Zeng K., Fu Y. X., Shi S., Wu C. I. (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174: 1431-9
36. Zeng K., Shi S., Wu C. I. (2007) Compound tests for the detection of hitchhiking under positive selection. Mol Biol Evol 24: 1898-908
37. Zhao Q., Thuillet A. C., Uhlmann N. K., Weber A., Rafalski J. A., Allen S. M., Tingey S., Doebley J. (2008) The role of regulatory genes during maize domestication: evidence from nucleotide polymorphism and gene expression. Genetics 178: 2133-43
38. 林栲，李海鹏. (2009) DNA 水平上检测正向选择方法的研究进展. Herditas 31: 896-902
39. 周琦，王文. (2004) DNA 水平自然选择作用的检验. Zoological Research 25: 73-80

附录： 生物信息学主要英文术语及释义

Abstract Syntax Notation (ASN.l)（**NCBI**发展的许多程序，如显示蛋白质三维结构的**Cn3D**等所使用的内部格式）
A language that is used to describe structured data types formally, Within bioinformatits,it has been used by the National Center for Biotechnology Information to encode sequences, maps, taxonomic information, molecular structures, and biographical information in such a way that it can be easily accessed and exchanged by computer software.
Accession number（记录号）
A unique identifier that is assigned to a single database entry for a DNA or protein sequence.
Affine gap penalty（一种设置空位罚分策略）
A gap penalty score that is a linear function of gap length, consisting of a gap opening penalty and a gap extension penalty multiplied by the length of the gap. Using this penalty scheme greatly enhances the performance of dynamic programming methods for sequence alignment. See also Gap penalty.
Algorithm（算法）
A systematic procedure for solving a problem in a finite number of steps, typically involving a repetition of operations. Once specified, an algorithm can be written in a computer language and run as a program.
Alignment（联配/比对/联配）
Refers to the procedure of comparing two or more sequences by looking for a series of individual characters or character patterns that are in the same order in the sequences. Of the two types of alignment, local and global, a local alignment is generally the most useful. See also Local and Global alignments.
Alignment score（联配/比对/联配值）
An algorithmically computed score based on the number of matches, substitutions, insertions, and deletions (gaps) within an alignment. Scores for matches and substitutions Are derived from a scoring matrix such as the BLOSUM and PAM matrices for proteins, and aftine gap penalties suitable for the matrix are chosen. Alignment scores are in log odds units, often bit units (log to the base 2). Higher scores denote better alignments. See also Similarity score, Distance in sequence analysis.
Alphabet（字母表）
The total number of symbols in a sequence-4 for DNA sequences and 20 for protein sequences.
Annotation（注释）
The prediction of genes in a genome, including the location of protein-encoding genes, the sequence of the encoded proteins, any significant

matches to other Proteins of known function, and the location of RNA-encoding genes. Predictions are based on gene models; e.g., hidden Markov models of introns and exons in proteins encoding genes, and models of secondary structure in RNA.

Anonymous FTP（匿名**FTP**）

When a FTP service allows anyone to log in, it is said to provide anonymous FTP ser-vice. A user can log in to an anonymous FTP server by typing anonymous as the user name and his E-mail address as a password. Most Web browsers now negotiate anonymous FTP logon without asking the user for a user name and password. See also FTP.

ASCII

The American Standard Code for Information Interchange (ASCII) encodes unaccented letters a-z, A-Z, the numbers O-9, most punctuation marks, space, and a set of control characters such as carriage return and tab. ASCII specifies 128 characters that are mapped to the values O-127. ASCII tiles are commonly called plain text, meaning that they only encode text without extra markup.

BAC clone（细菌人工染色体克隆）

Bacterial artificial chromosome vector carrying a genomic DNA insert, typically 100–200 kb. Most of the large-insert clones sequenced in the project were BAC clones.

Back-propagation（反向传输）

When training feed-forward neural networks, a back-propagation algorithm can be used to modify the network weights. After each training input pattern is fed through the network, the network's output is compared with the desired output and the amount of error is calculated. This error is back-propagated through the network by using an error function to correct the network weights. See also Feed-forward neural network.

Baum-Welch algorithm（**Baum-Welch**算法）

An expectation maximization algorithm that is used to train hidden Markov models.

Baye's rule（贝叶斯法则）

Forms the basis of conditional probability by calculating the likelihood of an event occurring based on the history of the event and relevant background information. In terms of two parameters A and B, the theorem is stated in an equation: The condition-al probability of *A,* given B, P(AIB), is equal to the probability of A, P(A), times the conditional probability of B, given A, P(BIA), divided by the probability of B, *P(B). P(A)* is the historical or prior distribution value of *A, P(BIA)* is a new prediction for B for a particular value of *A,* and *P(B)* is the sum of the newly predicted values for B. *P(AIB)* is a posterior probability, representing a new prediction for *A* given the prior knowledge of *A* and the newly discovered relationships between *A* and B.

Bayesian analysis（贝叶斯分析）

A statistical procedure used to estimate parameters of an underlying

distribution based on an observed distribution. See also Baye’s rule.

Biochips（生物芯片）

Miniaturized arrays of large numbers of molecular substrates, often oligonucleotides, in a defined pattern. They are also called DNA microarrays and microchips.

Bioinformatics （生物信息学）

The merger of biotechnology and information technology with the goal of revealing new insights and principles in biology. /The discipline of obtaining information about genomic or protein sequence data. This may involve similarity searches of databases, comparing your unidentified sequence to the sequences in a database, or making predictions about the sequence based on current knowledge of similar sequences. Databases are frequently made publically available through the Internet, or locally at your institution.

Bit score （二进制值/ Bit 值）

The value S' is derived from the raw alignment score S in which the statistical properties of the scoring system used have been taken into account. Because bit scores have been normalized with respect to the scoring system, they can be used to compare alignment scores from different searches.

Bit units

From information theory, a bit denotes the amount of information required to distinguish between two equally likely possibilities. The number of bits of information, AJ, required to convey a message that has A4 possibilities is log2 *M* = N bits.

BLAST （基本局部联配搜索工具，一种主要数据库搜索程序）

Basic Local Alignment Search Tool. A set of programs, used to perform fast similarity searches. Nucleotide sequences can be compared with nucleotide sequences in a database using BLASTN, for example. Complex statistics are applied to judge the significance of each match. Reported sequences may be homologous to, or related to the query sequence. The BLASTP program is used to search a protein database for a match against a query protein sequence. There are several other flavours of BLAST. BLAST2 is a newer release of BLAST. Allows for insertions or deletions in the sequences being aligned. Gapped alignments may be more biologically significant.

Block（蛋白质家族中保守区域的组块）

Conserved ungapped patterns approximately 3-60 amino acids in length in a set of related proteins.

BLOSUM matrices（模块替换矩阵，一种主要替换矩阵）

An alternative to PAM tables, BLOSUM tables were derived using local multiple alignments of more distantly related sequences than were used for the PAM matrix. These are used to assess the similarity of sequences when performing alignments.

Boltzmann distribution（**Boltzmann** 分布）

Describes the number of molecules that have energies above a certain level, based on the Boltzmann gas constant and the absolute temperature.

Boltzmann probability function(Boltzmann概率函数)
See Boltzmann distribution.
Bootstrap analysis
A method for testing how well a particular data set fits a model. For example, the validity of the branch arrangement in a predicted phylogenetic tree can be tested by resampling columns in a multiple sequence alignment to create many new alignments. The appearance of a particular branch in trees generated from these resampled sequences can then be measured. Alternatively, a sequence may be left out of an analysis to deter-mine how much the sequence influences the results of an analysis.
Branch length（分支长度）
In sequence analysis, the number of sequence changes along a particular branch of a phylogenetic tree.
CDS or cds （编码序列）
Coding sequence.
Chebyshe, d inequality
The probability that a random variable exceeds its mean is less than or equal to the square of 1 over the number of standard deviations from the mean.
Clone （克隆）
Population of identical cells or molecules (e.g. DNA), derived from a single ancestor.
Cloning Vector （克隆载体）
A molecule that carries a foreign gene into a host, and allows/facilitates the multiplication of that gene in a host. When sequencing a gene that has been cloned using a cloning vector (rather than by PCR), care should be taken not to include the cloning vector sequence when performing similarity searches. Plasmids, cosmids, phagemids, YACs and PACs are example types of cloning vectors.
Cluster analysis（聚类分析）
A method for grouping together a set of objects that are most similar from a larger group of related objects. The relationships are based on some criterion of similarity or difference. For sequences, a similarity or distance score or a statistical evaluation of those scores is used.
Cobbler
A single sequence that represents the most conserved regions in a multiple sequence alignment. The BLOCKS server uses the cobbler sequence to perform a database similarity search as a way to reach sequences that are more divergent than would be found using the single sequences in the alignment for searches.
Coding system (neural networks)
Regarding neural networks, a coding system needs to be designed for representing input and output. The level of success found when training the model will be partially dependent on the quality of the coding system chosen.
Codon usage

Analysis of the codons used in a particular gene or organism.

COG（直系同源簇）

Clusters of orthologous groups in a set of groups of related sequences in microorganism and yeast (S. *cerevisiae).* These groups are found by whole proteome comparisons and include orthologs and paralogs. See also Orthologs and Paralogs.

Comparative genomics（比较基因组学）

A comparison of gene numbers, gene locations, and biological functions of genes in the genomes of diverse organisms, one objective being to identify groups of genes that play a unique biological role in a particular organism.

Complexity (of an algorithm)（算法的复杂性）

Describes the number of steps required by the algorithm to solve a problem as a function of the amount of data; for example, the length of sequences to be aligned.

Conditional probability（条件概率）

The probability of a particular result (or of a particular value of a variable) given one or more events or conditions (or values of other variables).

Conservation （保守）

Changes at a specific position of an amino acid or (less commonly, DNA) sequence that preserve the physico-chemical properties of the original residue.

Consensus（一致序列）

A single sequence that represents, at each subsequent position, the variation found within corresponding columns of a multiple sequence alignment.

Context-free grammars

A recursive set of production rules for generating patterns of strings. These consist of a set of terminal characters that are used to create strings, a set of nonterminal symbols that correspond to rules and act as placeholders for patterns that can be generated using terminal characters, a set of rules for replacing nonterminal symbols with terminal characters, and a start symbol.

Contig （序列重叠群/拼接序列）

A set of clones that can be assembled into a linear order. A DNA sequence that overlaps with another contig. The full set of overlapping sequences (contigs) can be put together to obtain the sequence for a long region of DNA that cannot be sequenced in one run in a sequencing assay. Important in genetic mapping at the molecular level.

CORBA（国际对象管理协作组制定的使**OOP**对象与网络接口统一起来的一套跨计算机、操作系统、程序语言和网络的共同标准）

The Common Object Request Broker Architecture (CORBA) is an open industry standard for working with distributed objects, developed by the Object Management Group. CORBA allows the interconnection of objects and applications regardless of computer language, machine architecture, or geographic location of the computers.

Correlation coefficient（相关系数）

A numerical measure, falling between - **1** and **1,** of the degree of the linear relationship between two variables. A positive value indicates a direct relationship, a negative value indicates an inverse relationship, and the distance of the value away from zero indicates the strength of the relationship. A value near zero indicates no relationship between the variables.

Covariation (in sequences)（共变）

Coincident change at two or more sequence positions in related sequences that may influence the secondary structures of RNA or protein molecules.

Coverage (or depth) （覆盖率/厚度）

The average number of times a nucleotide is represented by a high-quality base in a collection of random raw sequence. Operationally, a 'high-quality base' is defined as one with an accuracy of at least 99% (corresponding to a PHRED score of at least 20).

Database（数据库）

A computerized storehouse of data that provides a standardized way for locating, adding, removing, and changing data. See also Object-oriented database, Relational database.

Dendogram

A form of a tree that lists the compared objects (e.g., sequences or genes in a microarray analysis) in a vertical order and joins related ones by levels of branches extending to one side of the list.

Depth （厚度）

See coverage

Dirichlet mixtures

Defined as the conjugational prior of a multinomial distribution. One use is for predicting the expected pattern of amino acid variation found in the match state of a hid-den Markov model (representing one column of a multiple sequence alignment of proteins), based on prior distributions found in conserved protein domains (blocks).

Distance in sequence analysis（序列距离）

The number of observed changes in an optimal alignment of two sequences, usually not counting gaps.

DNA Sequencing （**DNA** 测序）

The experimental process of determining the nucleotide sequence of a region of DNA. This is done by labelling each nucleotide (A, C, G or T) with either a radioactive or fluorescent marker which identifies it. There are several methods of applying this technology, each with their advantages and disadvantages. For more information, refer to a current text book. High throughput laboratories frequently use automated sequencers, which are capable of rapidly reading large numbers of templates. Sometimes, the sequences may be generated more quickly than they can be characterised.

Domain （功能域）

A discrete portion of a protein assumed to fold independently of the rest of the protein and possessing its own function.

Dot matrix（点标矩阵图）
Dot matrix diagrams provide a graphical method for comparing two sequences. One sequence is written horizontally across the top of the graph and the other along the left-hand side. Dots are placed within the graph at the intersection of the same letter appearing in both sequences. A series of diagonal lines in the graph indicate regions of alignment. The matrix may be filtered to reveal the most-alike regions by scoring a minimal threshold number of matches within a sequence window.
Draft genome sequence （基因组序列草图）
The sequence produced by combining the information from the individual sequenced clones (by creating merged sequence contigs and then employing linking information to create scaffolds) and positioning the sequence along the physical map of the chromosomes.
DUST （一种低复杂性区段过滤程序）
A program for filtering low complexity regions from nucleic acid sequences.
Dynamic programming（动态规划法）
A dynamic programming algorithm solves a problem by combining solutions to sub-problems that are computed once and saved in a table or matrix. Dynamic programming is typically used when a problem has many possible solutions and an optimal one needs to be found. This algorithm is used for producing sequence alignments, given a scoring system for sequence comparisons.
EMBL （欧洲分子生物学实验室，**EMBL** 数据库是主要公共核酸序列数据库之一）
European Molecular Biology Laboratories. Maintain the EMBL database, one of the major public sequence databases.
EMBnet （欧洲分子生物学网络）
European Molecular Biology Network: http://www.embnet.org/ was established in 1988, and provides services including local molecular databases and software for molecular biologists in Europe. There are several large outposts of EMBnet, including EXPASY.
Entropy（熵）
From information theory, a measure of the unpredictable nature of a set of possible elements. The higher the level of variation within the set, the higher the entropy.
Erdos and Renyi law
In a toss of a “fair” coin, the number of heads in a row that can be expected is the logarithm of the number of tosses to the base 2. The law may be generalized for more than two possible outcomes by changing the base of the logarithm to the number of out-comes. This law was used to analyze the number of matches and mismatches that can be expected between random sequences as a basis for scoring the statistical significance of a sequence alignment.
EST （表达序列标签的缩写）

See Expressed Sequence Tag

Expect value (E)（**E**值）

E value. The number of different alignents with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score. In a database similarity search, the probability that an alignment score as good as the one found between a query sequence and a database sequence would be found in as many comparisons between random sequences as was done to find the matching sequence. In other types of sequence analysis, E has a similar meaning.

Expectation maximization (sequence analysis)

An algorithm for locating similar sequence patterns in a set of sequences. A guessed alignment of the sequences is first used to generate an expected scoring matrix representing the distribution of sequence characters in each column of the alignment, this pattern is matched to each sequence, and the scoring matrix values are then updated to maximize the alignment of the matrix to the sequences. The procedure is repeated until there is no further improvement.

Exon （外显子）

Coding region of DNA. See CDS.

Expressed Sequence Tag (EST) （表达序列标签）

Randomly selected, partial cDNA sequence; represents it's corresponding mRNA. dbEST is a large database of ESTs at GenBank, NCBI.

FASTA （一种主要数据库搜索程序）

The first widely used algorithm for database similarity searching. The program looks for optimal local alignments by scanning the sequence for small matches called "words". Initially, the scores of segments in which there are multiple word hits are calculated ("init1"). Later the scores of several segments may be summed to generate an "initn" score. An optimized alignment that includes gaps is shown in the output as "opt". The sensitivity and speed of the search are inversely related and controlled by the "k-tup" variable which specifies the size of a "word". (Pearson and Lipman)

Extreme value distribution（极值分布）

Some measurements are found to follow a distribution that has a long tail which decays at high values much more slowly than that found in a normal distribution. This slow-falling type is called the extreme value distribution. The alignment scores between unrelated or random sequences are an example. These scores can reach very high values, particularly when a large number of comparisons are made, as in a database similarity search. The probability of a particular score may be accurately predicted by the extreme value distribution, which follows a double negative exponential function after Gumbel.

False negative（假阴性）

A negative data point collected in a data set that was incorrectly reported due to a failure of the test in avoiding negative results.

False positive （假阳性）
A positive data point collected in a data set that was incorrectly reported due to a failure of the test. If the test had correctly measured the data point, the data would have been recorded as negative.

Feed-forward neural network （反向传输神经网络）
Organizes nodes into sequence layers in which the nodes in each layer are fully connected with the nodes in the next layer, except for the final output layer. Input is fed from the input layer through the layers in sequence in a “feed-forward” direction, resulting in output at the final layer. See also Neural network.

Filtering (window size)
During pair-wise sequence alignment using the dot matrix method, random matches can be filtered out by using a sliding window to compare the two sequences. Rather than comparing a single sequence position at a time, a window of adjacent positions in the two sequences is compared and a dot, indicating a match, is generated only if a certain minimal number of matches occur.

Filtering （过滤）
Also known as Masking. The process of hiding regions of (nucleic acid or amino acid) sequence having characteristics that frequently lead to spurious high scores. See SEG and DUST.

Finished sequence（完成序列）
Complete sequence of a clone or genome, with an accuracy of at least 99.99% and no gaps.

Fourier analysis
Studies the approximations and decomposition of functions using trigonometric polynomials.

Format (file)（格式）
Different programs require that information be specified to them in a formal manner, using particular keywords and ordering. This specification is a file format.

Forward-backward algorithm
Used to train a hidden Markov model by aligning the model with training sequences. The algorithm then refines the model to reduce the error when fitted to the given data using a gradient descent approach.

FTP (File Transfer Protocol)（文件传输协议）
Allows a person to transfer files from one computer to another across a network using an FTP-capable client program. The FTP client program can only communicate with machines that run an FTP server. The server, in turn, will make a specific portion of its tile system available for FTP access, providing that the client is able to supply a recognized user name and password to the server.

Full shotgun clone （鸟枪法克隆）
A large-insert clone for which full shotgun sequence has been produced.

Functional genomics（功能基因组学）
Assessment of the function of genes identified by between-genome comparisons. The function of a newly identified gene is tested by introducing mutations into the gene and then examining the resultant mutant organism for an altered phenotype.
gap （空位/间隙/缺口）
A space introduced into an alignment to compensate for insertions and deletions in one sequence relative to another. To prevent the accumulation of too many gaps in an alignment, introduction of a gap causes the deduction of a fixed amount (the gap score) from the alignment score. Extension of the gap to encompass additional nucleotides or amino acid is also penalized in the scoring of an alignment.
Gap penalty（空位罚分）
A numeric score used in sequence alignment programs to penalize the presence of gaps within an alignment. The value of a gap penalty affects how often gaps appear in alignments produced by the algorithm. Most alignment programs suggest gap penalties that are appropriate for particular scoring matrices.
Genetic algorithm（遗传算法）
A kind of search algorithm that was inspired by the principles of evolution. A population of initial solutions is encoded and the algorithm searches through these by applying a pre-defined fitness measurement to each solution, selecting those with the highest fitness for reproduction. New solutions can be generated during this phase by crossover and mutation operations, defined in the encoded solutions.
Genetic map （遗传图谱）
A genome map in which polymorphic loci are positioned relative to one another on the basis of the frequency with which they recombine during meiosis. The unit of distance is centimorgans (cM), denoting a 1% chance of recombination.
Genome（基因组）
The genetic material of an organism, contained in one haploid set of chromosomes.
Gibbs sampling method
An algorithm for finding conserved patterns within a set of related sequences. A guessed alignment of all but one sequence is made and used to generate a scoring matrix that represents the alignment. The matrix is then matched to the left-out sequence, and a probable location of the corresponding pattern is found. This prediction is then input into a new alignment and another scoring matrix is produced and tested on a new left-out sequence. The process is repeated until there is no further improvement in the matrix.
Global alignment（整体联配）
Attempts to match as many characters as possible, from end to end, in a set of two or

more sequences.

Gopher (一个文档发布系统，允许检索和显示文本文件)

Graph theory（图论）

A branch of mathematics which deals with problems that involve a graph or network structure. A graph is defined by a set of nodes (or points) and a set of arcs (lines or edges) joining the nodes. In sequence and genome analysis, graph theory is used for sequence alignments and clustering alike genes.

GSS（基因综述序列）

Genome survey sequence.

GUI（图形用户界面）

Graphical user interface.

H （相对熵值）

H is the relative entropy of the target and background residue frequencies. (Karlin and Altschul, 1990). H can be thought of as a measure of the average information (in bits) available per position that distinguishes an alignment from chance. At high values of H, short alignments can be distinguished by chance, whereas at lower H values, a longer alignment may be necessary. (Altschul, 1991)

Half-bits

Some scoring matrices are in half-bit units. These units are logarithms to the base 2 of odds scores times 2.

Heuristic（启发式方法）

A procedure that progresses along empirical lines by using rules of thumb to reach a solution. The solution is not guaranteed to be optimal.

Hexadecimal system（**16**制系统）

The base 16 counting system that uses the digits O-9 followed by the letters A-F.

HGMP （人类基因组图谱计划）

Human Genome Mapping Project.

Hidden Markov Model (HMM)（隐马尔可夫模型）

In sequence analysis, a HMM is usually a probabilistic model of a multiple sequence alignment, but can also be a model of periodic patterns in a single sequence, representing, for example, patterns found in the exons of a gene. In a model of multiple sequence alignments, each column of symbols in the alignment is represented by a frequency distribution of the symbols called a state, and insertions and deletions by other states. One then moves through the model along a particular path from state to state trying to match a given sequence. The next matching symbol is chosen from each state, recording its probability (frequency) and also the probability of going to that particular state from a previous one (the transition probability). State and transition probabilities are then multiplied to obtain a probability of the given sequence. Generally speaking, a HMM is a statistical model for an ordered sequence of symbols, acting as a stochastic state machine that generates a symbol each time a transition is made from one state to the next. Transitions between

states are specified by transition probabilities.

Hidden layer（隐藏层）

An inner layer within a neural network that receives its input and sends its output to other layers within the network. One function of the hidden layer is to detect covariation within the input data, such as patterns of amino acid covariation that are associated with a particular type of secondary structure in proteins.

Hierarchical clustering（分级聚类）

The clustering or grouping of objects based on some single criterion of similarity or difference.An example is the clustering of genes in a microarray experiment based on the correlation between their expression patterns. The distance method used in phylogenetic analysis is another example.

Hill climbing

A nonoptimal search algorithm that selects the singular best possible solution at a given state or step. The solution may result in a locally best solution that is not a globally best solution.

Homology（同源性）

A similar component in two organisms (e.g., genes with strongly similar sequences) that can be attributed to a common ancestor of the two organisms during evolution.

Horizontal transfer（水平转移）

The transfer of genetic material between two distinct species that do not ordinarily exchange genetic material. The transferred DNA becomes established in the recipient genome and can be detected by a novel phylogenetic history and codon content com-pared to the rest of the genome.

HSP （高比值片段对）

High-scoring segment pair. Local alignments with no gaps that achieve one of the top alignment scores in a given search.

HTGS/HGT（高通量基因组序列）

High-throughout genome sequences

HTML（超文本标识语言）

The Hyper-Text Markup Language (HTML) provides a structural description of a document using a specified tag set. HTML currently serves as the Internet lingua franca for describing hypertext Web page documents.

Hyperplane

A generalization of the two-dimensional plane to N dimensions.

Hypercube

A generalization of the three-dimensional cube to N dimensions.

Identity （相同性/相同率）

The extent to which two (nucleotide or amino acid) sequences are invariant.

Indel（插入或删除的缩略语）

An insertion or deletion in a sequence alignment.

Information content (of a scoring matrix)

A representation of the degree of sequence conservation in a column of a

scoring matrix representing an alignment of related sequences. It is also the number of questions that must be asked to match the column to a position in a test sequence. For bases, the max-imum possible number is 2, and for proteins, 4.32 (logarithm to the base 2 of the number of possible sequence characters).

Information theory（信息理论）

A branch of mathematics that measures information in terms of bits, the minimal amount of structural complexity needed to encode a given piece of information.

Input layer（输入层）

The initial layer in a feed-forward neural net. This layer encodes input information that will be fed through the network model.

Interface definition language

Used to define an interface to an object model in a programming language neutral form, where an interface is an abstraction of a service defined only by the operations that can be performed on it.

Internet（因特网）

The network infrastructure, consisting of cables interconnected by routers, that pro-vides global connectivity for individual computers and private networks of computers. A second sense of the word internet is the collective computer resources available over this global network.

Interpolated Markov model

A type of Markov model of sequences that examines sequences for patterns of variable length in order to discriminate best between genes and non-gene sequences.

Intranet（内部网）

Intron （内含子）

Non-coding region of DNA.

Iterative（反复的/迭代的）

A sequence of operations in a procedure that is performed repeatedly.

Java（一种由 **SUN Microsystem** 开发的编程语言）

K （**BLAST** 程序的一个统计参数）

A statistical parameter used in calculating BLAST scores that can be thought of as a natural scale for search space size. The value K is used in converting a raw score (S) to a bit score (S').

K-tuple（字/字长）

Identical short stretches of sequences, also called words.

lambda （λ，**BLAST** 程序的一个统计参数）

A statistical parameter used in calculating BLAST scores that can be thought of as a natural scale for scoring system. The value lambda is used in converting a raw score (S) to a bit score (S').

LAN（局域网）

Local area network.

Likelihood（似然性）

The hypothetical probability that an event which has already occurred would yield a specific outcome. Unlike probability, which refers to future events, likelihood refers to past events.

Linear discriminant analysis

An analysis in which a straight line is located on a graph between two sets of data pointsin a location that best separates the data points into two groups.

Local alignment（局部联配）

Attempts to align regions of sequences with the highest density of matches. In doing so, one or more islands of subalignments are created in the aligned sequences.

Log odds score（概率对数值）

The logarithm of an odds score. See also Odds score.

Low Complexity Region (LCR) （低复杂性区段）

Regions of biased composition including homopolymeric runs, short-period repeats, and more subtle overrepresentation of one or a few residues. The SEG program is used to mask or filter LCRs in amino acid queries. The DUST program is used to mask or filter LCRs in nucleic acid queries.

Machine learning（机器学习）

The training of a computational model of a process or classification scheme to distinguish between alternative possibilities.

Markov chain（马尔可夫链）

Describes a process that can be in one of a number of states at any given time. The Markov chain is defined by probabilities for each transition occurring; that is, probabilities of the occurrence of state sj given that the current state is sp Substitutions in nucleic acid and protein sequences are generally assumed to follow a Markov chain in that each site changes independently of the previous history of the site. With this model, the number and types of substitutions observed over a relatively short period of evolutionary time can be extrapolated to longer periods of time. In performing sequence alignments and calculating the statistical significance of alignment scores, sequences are assumed to be Markov chains in which the choice of one sequence position is not influenced by another.

Masking （过滤）

Also known as Filtering. The removal of repeated or low complexity regions from a sequence in order to improve the sensitivity of sequence similarity searches performed with that sequence.

Maximum likelihood (phylogeny, alignment)（最大似然法）

The most likely outcome (tree or alignment), given a probabilistic model of evolutionary change in DNA sequences.

Maximum parsimony（最大简约法）

The minimum number of evolutionary steps required to generate the observed variation in a set of sequences, as found by comparison of the number of steps in all possible phylogenetic trees.

Method of moments

The mean or expected value of a variable is the first moment of the values of the variable around the mean, defined as that number from which the sum of deviations to all values is zero. The standard deviation is the second moment of the values about the mean, and so on.

Minimum spanning tree

Given a set of related objects classified by some similarity or difference score, the mini-mum spanning tree joins the most-alike objects on adjacent outer branches of a tree and then sequentially joins less-alike objects by more inward branches. The tree branch lengths are calculated by the same neighbor-joining algorithm that is used to build phylogenetic trees of sequences from a distance matrix. The sum of the resulting branch lengths between each pair of objects will be approximately that found by the classification scheme.

MMDB （分子建模数据库）

Molecular Modelling Database. A taxonomy assigned database of PDB (see PDB) files, and related information.

Molecular clock hypothesis（分子钟假设）

The hypothesis that sequences change at the same rate in the branches of an evolutionary
tree.

Monte Carlo（蒙特卡罗法）

A method that samples possible solutions to a complex problem as a way to estimate a more general solution.

Motif （模序）

A short conserved region in a protein sequence. Motifs are frequently highly conserved parts of domains.

Multiple Sequence Alignment （多序列联配）

An alignment of three or more sequences with gaps inserted in the sequences such that residues with common structural positions and/or ancestral residues are aligned in the same column. Clustal W is one of the most widely used multiple sequence alignment programs

Mutation data matrix（突变数据矩阵，即**PAM**矩阵）

A scoring matrix compiled from the observation of point mutations between aligned sequences. Also refers to a Dayhoff PAM matrix in which the scores are given as log odds scores.

N50 length （**N50** 长度，即覆盖 **50%**所有核苷酸的最大序列重叠群长度）

A measure of the contig length (or scaffold length) containing a 'typical' nucleotide. Specifically, it is the maximum length L such that 50% of all nucleotides lie in contigs (or scaffolds) of size at least L.

Nats (natural logarithm)

A number expressed in units of the natural logarithm.

NCBI （美国国家生物技术信息中心）

National Center for Biotechnology Information (USA). Created by the United States Congress in 1988, to develop information systems to support the

biological research community.

Needleman-Wunsch algorithm（**Needleman-Wunsch**算法）

Uses dynamic programming to find global alignments between sequences.

Neighbor-joining method（邻接法）

Clusters together alike pairs within a group of related objects (e.g., genes with similar sequences) to create a tree whose branches reflect the degrees of difference among the objects.

Neural network（神经网络）

From artificial intelligence algorithms, techniques that involve a set of many simple units that hold symbolic data, which are interconnected by a network of links associated with numeric weights. Units operate only on their symbolic data and on the inputs that they receive through their connections. Most neural networks use a training algorithm (see Back-propagation) to adjust connection weights, allowing the network to learn associations between various input and output patterns. See also Feed-forward neural network.

NIH （美国国家卫生研究院）

National Institutes of Health (USA).

Noise（噪音）

In sequence analysis, a small amount of randomly generated variation in sequences that is added to a model of the sequences; e.g., a hidden Markov model or scoring matrix, in order to avoid the model overfitting the sequences. See also Overfitting.

Normal distribution（正态分布）

The distribution found for many types of data such as body weight, size, and exam scores. The distribution is a bell-shaped curve that is described by a mean and standard deviation of the mean. Local sequence alignment scores between unrelated or random sequences do not follow this distribution but instead the extreme value distribution which has a much extended tail for higher scores. See also Extreme value distribution.

Object Management Group (OMG)（国际对象管理协作组）

A not-for-profit corporation that was formed to promote component-based software by introducing standardized object software. The OMG establishes industry guidelines and detailed object management specifications in order to provide a common framework for application development. Within OMG is a Life Sciences Research group, a consortium representing pharmaceutical companies, academic institutions, software vendors, and hardware vendors who are working together to improve communication and inter-operability among computational resources in life sciences research. See CORBA.

Object-oriented database（面向对象数据库）

Unlike relational databases (see entry), which use a tabular structure, object-oriented databases attempt to model the structure of a given data set as closely as possible. In doing so, object-oriented databases tend to reduce the appearance of duplicated data and the complexity of query structure often found in relational databases.

Odds score（概率/几率值）
The ratio of the likelihoods of two events or outcomes. In sequence alignments and scoring matrices, the odds score for matching two sequence characters is the ratio of the frequency with which the characters are aligned in related sequences divided by the frequency with which those same two characters align by chance alone, given the frequency of occurrence of each in the sequences. Odds scores for a set of individually aligned positions are obtained by multiplying the odds scores for each position. Odds scores are often converted to logarithms to create log odds scores that can be added to obtain the log odds score of a sequence alignment.
OMIM （一种人类遗传疾病数据库）
Online Mendelian Inheritance in Man. Database of genetic diseases with references to molecular medicine, cell biology, biochemistry and clinical details of the diseases.
Optimal alignment（最佳联配）
The highest-scoring alignment found by an algorithm capable of producing multiple solutions. This is the best possible alignment that can be found, given any parameters supplied by the user to the sequence alignment program.
ORF （开放阅读框）
Open Reading Frame. A series of codons (base triplets) which can be translated into a protein. There are six potential reading frames of an unidentifed sequence; TBLASTN (see BLAST) transalates a nucleotide sequence in all six reading frames, into a protein, then attempts to align the results to sequeneces in a protein database, returning the results as a nucleotide sequence. The most likely reading frame can be identified using on-line software (e.g. ORF Finder).
Orthologous（直系同源）
Homologous sequences in different species that arose from a common ancestral gene during speciation; may or may not be responsible for a similar function. A pair of genes found in two species are orthologous when the encoded proteins are 60-80% identical in an alignment. The proteins almost certainly have the same three-dimensional structure, domain structure, and biological function, and the encoding genes have originated from a common ancestor gene at an earlier evolutionary time. Two orthologs 1 and II in genomes A and B, respectively, may be identified when the complete genomes of two species are available: (1) in a database similarity search of all of the proteome of B using I as a query, II is the best hit found, and (2) I is the best hit when 11 is used as a query of the proteome of B. The best hit is the database sequence with the highest expect value (E). Orthology is also predicted by a very close phylogenetic relationship between sequences or by a cluster analysis. Compare to Paralogs. See also Cluster analysis.
Output layer（输出层）
The final layer of a neural network in which signals from lower levels in the network are input into output states where they are weighted and summed to

give an outpu t signal. For example, the output signal might be the prediction of one type of protein secondary structure for the central amino acid in a sequence window.

Overfitting

Can occur when using a learning algorithm to train a model such as a neural net or hid-den Markov model. Overfitting refers to the model becoming too highly representative of the training data and thus no longer representative of the overall range of data that is supposed to be modeled.

P value （**P** 值/概率值）

The probability of an alignment occurring with the score in question or better. The p value is calculated by relating the observed alignment score, S, to the expected distribution of HSP scores from comparisons of random sequences of the same length and composition as the query to the database. The most highly significant P values will be those close to 0. P values and E values are different ways of representing the significance of the alignment.

Pair-wise sequence alignment（双序列联配）

An alignment performed between two sequences.

PAM （可接受突变百分率/可以观察到的突变百分率，它可作为一种进化时间单位）

Percent Accepted Mutation. A unit introduced by Dayhoff et al. to quantify the amount of evolutionary change in a protein sequence. 1.0 PAM unit, is the amount of evolution which will change, on average, 1% of amino acids in a protein sequence. A PAM(x) substitution matrix is a look-up table in which scores for each amino acid substitution have been calculated based on the frequency of that substitution in closely related proteins that have experienced a certain amount (x) of evolutionary divergence.

Paralogous （旁系同源）

Homologous sequences within a single species that arose by gene duplication. Genes that are related through gene duplication events. These events may lead to the production of a family of related proteins with similar biological functions within a species. Paralogous gene families within a species are identified by using an individual protein as a query in a database similarity search of the entireproteome of an organism. The process is repeated for the entire proteome and the resulting sets of related proteins are then searched for clusters that are most likely to have a conserved domain structure and should represent a paralogous gene family.

Parametric sequence alignment

An algorithm that finds a range of possible alignments based on varying the parameters of the scoring system for matches, mismatches, and gap penalties. An example is the Bayes block aligner.

PDB （主要蛋白质结构数据库之一）

Brookhaven Protein Data Bank. A database and format of files which describe the 3D structure of a protein or nucleic acid, as determined by X-ray crystallography or nuclear magnetic resonance (NMR) imaging. The

molecules described by the files are usually viewed locally by dedicated software, but can sometimes be visualised on the world wide web.

Pearson correlation coefficent（**Pearson**相关系数）

A measure of the correlation between two variables that reflects the degree to which the two variables are related. For example, the coefficient is used as a measure of similarity of gene expression in a microarray experiment. See also Correlation coefficient. Percent identity The percentage of the columns in an alignment of two sequences that includes identical amino acids. Columns in the alignment that include gaps are not scored in the calculation.

Percent similarity（相似百分率）

The percentage of the columns in an alignment of two sequences that includes either identical amino acids or amino acids that are frequently found substituted for each other in sequences of related proteins (conservative substitutions). These substitutions may be found in an amino acid substitution matrix such as the Dayhoff PAM and Henikoff BLOSUM matrices. Columns in the alignment that include gaps are not scored in the calculation.

Perceptron（感知器，模拟人类视神经控制系统的图形识别机**)**

A neural network in which input and output states are directly connected without intervening hidden layers.

PHRED （一种广泛应用的原始序列分析程序，可以对序列的各个碱基进行识别和质量评价）

A widely used computer program that analyses raw sequence to produce a 'base call' with an associated 'quality score' for each position in the sequence. A PHRED quality score of X corresponds to an error probability of approximately 10-X/10. Thus, a PHRED quality score of 30 corresponds to 99.9% accuracy for the base call in the raw read.

PHRAP （一种广泛应用的原始序列组装程序）

A widely used computer program that assembles raw sequence into sequence contigs and assigns to each position in the sequence an associated 'quality score', on the basis of the PHRED scores of the raw sequence reads. A PHRAP quality score of X corresponds to an error probability of approximately 10-X/10. Thus, a PHRAP quality score of 30 corresponds to 99.9% accuracy for a base in the assembled sequence.

Phylogenetic studies（系统发育研究）

PIR （主要蛋白质序列数据库之一，翻译自 **GenBank**）

A database of translated GenBank nucleotide sequences. PIR is a redundant (see Redundancy) protein sequence database. The database is divided into four categories:

PIR1 - Classified and annotated.

PIR2 - Annotated.

PIR3 - Unverified.

PIR4 - Unencoded or untranslated.

Poisson distribution（帕松分布）

Used to predict the occurrence of infrequent events over a long period of time

or when there are a large number of trials. In sequence analysis, it is used to calculate the chance that one pair of a large number of pairs of unrelated sequences may give a high local alignment score.

Position-specific scoring matrix (PSSM)（特定位点记分矩阵，**PSI-BLAST** 等搜索程序使用）

The PSSM gives the log-odds score for finding a particular matching amino acid in a target sequence. Represents the variation found in the columns of an alignment of a set of related sequences. Each subsequent matrix column corresponds to the next column in the alignment and each row corresponds to a particular sequence character (one of four bases in DNA sequences or 20 amino acids in protein sequences). Matrix values are log odds scores obtained by dividing the counts of the residue in the alignment, dividing by the expected number of counts based on sequence composition, and converting the ratio to a log score. The matrix is moved along sequences to find similar regions by adding the matching log odds scores and looking for high values. There is no allowance for gaps. Also called a weight matrix or scoring matrix.

Posterior (Bayesian analysis)

A conditional probability based on prior knowledge and newly evaluated relationships among variables using Bayes rule. See also Bayes rule.

Prior (Bayesian analysis)

The expected distribution of a variable based on previous data.

Profile（分布型）

A matrix representation of a conserved region in a multiple sequence alignment that allows for gaps in the alignment. The rows include scores for matching sequential columns of the alignment to a test sequence. The columns include substitution scores for amino acids and gap penalties. See also PSSM.

Profile hidden Markov model（分布型隐马尔可夫模型）

A hidden Markov model of a conserved region in a multiple sequence alignment that includes gaps and may be used to search new sequences for similarity to the aligned sequences.

Proteome（蛋白质组）

The entire collection of proteins that are encoded by the genome of an organism. Initially the proteome is estimated by gene prediction and annotation methods but eventually will be revised as more information on the sequence of the expressed genes is obtained.

Proteomics （蛋白质组学）

Systematic analysis of protein expression of normal and diseased tissues that involves the separation, identification and characterization of all of the proteins in an organism.

Pseudocounts

Small number of counts that is added to the columns of a scoring matrix to increase the variability either to avoid zero counts or to add more variation than was found in the sequences used to produce the matrix.

PSI-BLAST （**BLAST** 系列程序之一）
Position-Specific Iterative BLAST. An iterative search using the BLAST algorithm. A profile is built after the initial search, which is then used in subsequent searches. The process may be repeated, if desired with new sequences found in each cycle used to refine the profile. Details can be found in this discussion of PSI-BLAST. (Altschul et al.)

PSSM （特定位点记分矩阵）
See position-specific scoring matrix and profile.

Public sequence databases （公共序列数据库，指 **GenBank**、**EMBL** 和 **DDBJ**）
The three coordinated international sequence databases: GenBank, the EMBL data library and DDBJ.

Q20 (Quality score 20)
A quality score of > or = 20 indicates that there is less than a 1 in 100 chance that the base call is incorrect. These are consequently high-quality bases. Specifically, the quality value "q" assigned to a basecall is defined as:
q = -10 x log10(p)
where p is the estimated error probability for that basecall. Note that high quality values correspond to low error probabilities, and conversely.

Quality trimming
This is an algorithm which uses a sliding window of 50 bases and trims from the 5' end of the read followed by the 3' end. With each window, the number of low quality (10 or less) bases is determined. If more than 5 bases are below the threshold quality, the window is incremented by one base and the process is repeated. When the low quality test fails, the position where it stopped is recorded. The parameters for window length low quality threshold and number of low quality bases tolerated are fixed. The positions of the 5' and 3' boundaries of the quality region are noted in the plot of quality values presented in the" Chromatogram Details" report.

Query （待查序列/搜索序列）
The input sequence (or other type of search term) with which all of the entries in a database are to be compared.

Radiation hybrid (RH) map （辐射杂交图谱）
A genome map in which STSs are positioned relative to one another on the basis of the frequency with which they are separated by radiation-induced breaks. The frequency is assayed by analysing a panel of human–hamster hybrid cell lines, each produced by lethally irradiating human cells and fusing them with recipient hamster cells such that each carries a collection of human chromosomal fragments. The unit of distance is centirays (cR), denoting a 1% chance of a break occuring between two loci

Raw Score （初值，指最初得到的联配值 **S**）
The score of an alignment, S, calculated as the sum of substitution and gap scores. Substitution scores are given by a look-up table (see PAM, BLOSUM). Gap scores are typically calculated as the sum of G, the gap opening penalty

and L, the gap extension penalty. For a gap of length n, the gap cost would be G+Ln. The choice of gap costs, G and L is empirical, but it is customary to choose a high value for G (10-15)and a low value for L (1-2).

Raw sequence （原始序列/读胶序列）

Individual unassembled sequence reads, produced by sequencing of clones containing DNA inserts.

Receiver operator characteristic

The receiver operator characteristic (ROC) curve describes the probability that a test will correctly declare the condition present against the probability that the test will declare the condition present when actually absent. This is shown through a graph of the tesls sensitivity against one minus the test specificity for different possible threshold values.

Redundancy （冗余）

The presence of more than one identical item represents redundancy. In bioinformatics, the term is used with reference to the sequences in a sequence database. If a database is described as being redundant, more than one identical (redundant) sequence may be found. If the database is said to be non-redundant (nr), the database managers have attempted to reduce the redundancy. The term is ambiguous with reference to genetics, and as such, the degree of non-redundancy varies according to the database manager's interpretation of the term. One can argue whether or not two alleles of a locus defines the limit of redundancy, or whether the same locus in different, closely related organisms constitutes redundency. Non-redundant databases are, in some ways, superior, but are less complete. These factors should be taken into consideration when selecting a database to search.

Regular expressions

This computational tool provides a method for expressing the variations found in a set of related sequences including a range of choices at one position, insertions, repeats, and so on. For example, these expressions are used to characterize variations found in protein domains in the PROSITE catalog.

Regularization

A set of techniques for reducing data overfitting when training a model. See also Overfitting.

Relational database（关系数据库）

Organizes information into tables where each column represents the fields of informa-tion that can be stored in a single record. Each row in the table corresponds to a single record. A single database can have many tables and a query language is used to access the data. See also Object-oriented database.

Scaffold （支架，由序列重叠群拼接而成）

The result of connecting contigs by linking information from paired-end reads from plasmids, paired-end reads from BACs, known messenger RNAs or other sources. The contigs in a scaffold are ordered and oriented with respect to one another.

Scoring matrix（记分矩阵）
See Position-specific scoring matrix.
SEG （一种蛋白质程序低复杂性区段过滤程序）
A program for filtering low complexity regions in amino acid sequences. Residues that have been masked are represented as "X" in an alignment. SEG filtering is performed by default in the blastp subroutine of BLAST 2.0. (Wootton and Federhen)
Selectivity (in database similarity searches)（数据库相似性搜索的选择准确性）
The ability of a search method to locate members of a protein family without making a false-positive classification of members of other families.
Sensitivity (in database similarity searches)（数据库相似性搜索的灵敏性）
The ability of a search method to locate as many members of a protein family as possi-ble, including distant members of limited sequence similarity.
Sequence Tagged Site （序列标签位点）
Short cDNA sequences of regions that have been physically mapped. STSs provide unique landmarks, or identifiers, throughout the genome. Useful as a framework for further sequencing.
Significance（显著水平）
A significant result is one that has not simply occurred by chance, and therefore is prob-ably true. Significance levels show how likely a result is due to chance, expressed as a probability. In sequence analysis, the significance of an alignment score may be calcu-lated as the chance that such a score would be found between random or unrelated sequences. See Expect value.
Similarity score (sequence alignment) （相似性值）
Similarity means the extent to which nucleotide or protein sequences are related. The extent of similarity between two sequences can be based on percent sequence identity and/or conservation. In BLAST similarity refers to a positive matrix score. The sum of the number of identical matches and conservative (high scoring) substitu-tions in a sequence alignment divided by the total number of aligned sequence charac-ters. Gaps are usually ignored.
Simulated annealing
A search algorithm that attempts to solve the problem of finding global extrema. The algorithm was inspired by the physical cooling process of metals and the freezing process in liquids where atoms slow down in movement and line up to form a crystal. The algorithm traverses the energy levels of a function, always accepting energy levels that are smaller than previous ones, but sometimes accepting energy levels that are greater, according to the Boltzmann probability distribution.
Single-linkage cluster analysis
An analysis of a group of related objects, e.g., similar proteins in different genomes to identify both close and more distant relationships, represented on a tree or dendogram. The method joins the most closely related pairs by the neighbor-joining algorithm by representing these pairs as outer branches on

the tree. More distant objects are then pro-gressively added to lower tree branches. The method is also used to predict phylogenet-ic relationships by distance methods. See also Hierarchical clustering, Neighbor-joining method.

Smith-Waterman algorithm（**Smith-Waterman**算法）

Uses dynamic programming to find local alignments between sequences. The key fea-ture is that all negative scores calculated in the dynamic programming matrix are changed to zero in order to avoid extending poorly scoring alignments and to assist in identifying local alignments starting and stopping anywhere with the matrix.

SNP （单核苷酸多态性）

Single nucleotide polymorphism, or a single nucleotide position in the genome sequence for which two or more alternative alleles are present at appreciable frequency (traditionally, at least 1%) in the human population.

Space or time complexity（时间或空间复杂性）

An algorithms complexity is the maximum amount of computer memory or time required for the number of algorithmic steps to solve a problem.

Specificity (in database similarity searches)（数据库相似性搜索的特异性）

The ability of a search method to locate members of one protein family, including dis-tantly related members.

SSR （简单序列重复）

Simple sequence repeat, a sequence consisting largely of a tandem repeat of a specific k-mer (such as (CA)15). Many SSRs are polymorphic and have been widely used in genetic mapping.

Stochastic context-free grammar

A formal representation of groups of symbols in different parts of a sequence; i.e., not in the same context. An example is complementary regions in RNA that will form sec-ondary
structures. The stochastic feature introduces variability into such regions.

Stringency

Refers to the minimum number of matches required within a window. See also Filtering.

STS （序列标签位点的缩写）

See Sequence Tagged Site

Substitution （替换）

The presence of a non-identical amino acid at a given position in an alignment. If the aligned residues have similar physico-chemical properties the substitution is said to be "conservative".

Substitution Matrix （替换矩阵）

A substitution matrix containing values proportional to the probability that amino acid i mutates into amino acid j for all pairs of amino acids. such matrices are constructed by assembling a large and diverse sample of verified pairwise alignments of amino acids. If the sample is large enough to be statistically significant, the resulting matrices should reflect the true probabilities of mutations occuring through a period of evolution.

Sum of pairs method
Sums the substitution scores of all possible pair-wise combinations of sequence charac-ters in one column of a multiple sequence alignment.

SWISS-PROT （主要蛋白质序列数据库之一）
A non-redundant (See Redundancy) protein sequence database. Thoroughly annotated and cross referenced. A subdivision is TrEMBL.

Synteny
The presence of a set of homologous genes in the same order on two genomes.

Threading
In protein structure prediction, the aligning of the sequence of a protein of unknown structure with a known three-dimensional structure to determine whether the amino acid sequence is spatially and chemically compatible with that structure.

TrEMBL （蛋白质数据库之一，翻译自 **EMBL**）
A protein sequence database of Translated EMBL nucleotide sequences.

Uncertainty（不确定性）
From information theory, a logarithmic measure of the average number of choices that must be made for identification purposes. See also Information content.

Unified Modeling Language (UML)
A standard sanctioned by the Object Management Group that provides a formal nota-tion for describing object-oriented design.

UniGene （人类基因数据库之一）
Database of unique human genes, at NCBI. Entries are selected by near identical presence in GenBank and dbEST databases. The clusters of sequences produced are considered to represent a single gene.

Unitary Matrix （一元矩阵）
Also known as Identity Matrix. A scoring system in which only identical characters receive a positive score.

URL（统一资源定位符）
Uniform resource locator.

Viterbi algorithm
Calculates the optimal path of a sequence through a hidden Markov model of sequences using a dynamic programming algorithm.

Weight matrix
See Position-specific scoring matrix.

附录：核苷酸和氨基酸代码

（一）核苷酸代码

代码	核苷酸
A	Adenine 腺嘌呤
G	Guanine 鸟嘌呤
T	Thymine 胸腺嘧啶
C	Cytosine 胞嘧啶
U	Uracil 尿嘧啶

(二) IUB/IUPAC 代码

代码	碱基	说明
R	A 或 G	嘌呤
Y	T 或 C	嘧啶
W	A 或 T	弱键
S	C 或 G	强键
M	A 或 C	氨基
K	G 或 T	酮基
B	C, G 或 T	非 A
D	A, G 或 T	非 C
H	A, C 或 T	非 G
V	A, C 或 G	非 T
N	A, G, C 或 T	任意碱基

(三) 氨基酸代码

单字母代码	三字母代码		单字母代码	三字母代码	
A	Ala	Alanine 丙氨酸	M	Met	Methionine 甲硫氨酸 (蛋氨酸)
B	Asx	Asparagine 天冬酰胺 Aspartic acid 天冬氨酸	N	Asn	Asparagine 天冬酰胺
C	Cys	Cysteine 半胱氨酸	P	Pro	Proline 脯氨酸
D	Asp	Aspartic 天冬氨酸	Q	GIn	Glutamine 谷氨酰胺
E	Glu	Glutamic acid 谷氨酸	R	Arg	Arginine 精氨酸
F	Phe	Phenylalanine 苯丙氨酸	S	Ser	Serine 丝氨酸
G	Gly	Glycine 苷氨酸	T	Thr	Threonine 苏氨酸
H	His	Histidine 组氨酸	V	Val	Valine 缬氨酸
I	Ile	Isoleucine 异亮氨酸	W	Trp	Tryptophan 色氨酸
K	Lys	Lysine 赖氨酸	Y	Tyr	Tyrosine 脯氨酸
L	Leu	Leucme 亮氨酸	Z	Glz	Glutamine 谷氨酰胺 Glutamic acid 谷氨酸

(四)遗传密码

第一碱基	第二碱基				第三碱基
	U	C	A	G	
U	UUU Phe	UCU Ser	UAU tyr	UGU cys	U
	UUC Phe	UCC Ser	UAC tyr	UGC cys	C
	UUA Leu	UCA Ser	UAA -终止	UGA -终止	A
	UUG Leu	UCG Ser	UAG -终止	UAG -trp	G
C	CUU Leu	CCU Pro	CAU his	CGU arg	U
	CUC Leu	CCC Pro	CAC his	CGC arg	C
	CUA Leu	CCA Pro	CAA gln	CGA arg	A
	CUG Leu	CCG Pro	CAG gln	CGG arg	G
A	AUU ile	ACU thr	AAU asn	AGU ser	U
	AUC ile	ACC thr	AAC asn	AGC ser	C
	AUA ile	ACA thr	AAA lys	AGA arg	A
	AUG -Met	ACG thr	AAG lys	AGG arg	G
G	GUU Val	GCU ala	GAU asp	GGU gly	U
	GUC Val	GCC ala	GAC asp	GGC gly	C
	GUA Val	GCA ala	GAA glu	GGA gly	A
	GUG-起点	GCG ala	GAG glu	GGG gly	G

附录： 与核苷酸和蛋白质序列相关的特征关键词表

表 1 与核苷酸序列相关的特征关键词表

关键词	说明
allele	相关的个体或菌株含有相同基因的稳定的其它形式,该形式区别于这一位置的现有的序列(和或许其它序列)
attenuator	存在调节转录的终止的 DNA 区域,它控制了一些细菌操纵子的表达; (2)位于启动子和第一个结构基因之间，引起转录的部分终止的序列区段
C_region	免疫球蛋白轻和重链的恒定区,和 T-细胞受体 α，β，和 γ 链；根据特定的链可包括一个或多个外显子
CAAT_signal	CAAT 盒;位于可能参与 RNA 聚合酶结合的真核生物转录单位的起始点的 75bp 上游的保守序列的一部分；共有序列＝GG(C 或 T)CAATCT
CDS	编码序列;对应于蛋白质中的氨基酸序列的核苷酸的序列(位置包括终止密码子);特征包括氨基酸概念上的翻译
Conflict	在这一位点或区域,单独确定的“相同”序列有所不同
D-loop	置换环；线粒体 DNA 内的一个区域,其中 RNA 的短的序列与 DNA 的一条链配对,代替了这一区域的原始配对 DNA 链;也用于说明在 RecA 蛋白质催化的反应中，侵入的单链替代双链 DNA 的一条链的区域
D-segment	免疫球蛋白重链的多变区,和 T-细胞受体的 β 链
Enhancer	顺式-作用序列,它增强了(一些)真核生物启动子的作用,并能在任一方向和与启动子相关的任何位置处 (上游或下游)起作用
Exon	编码剪接 mRNA 部分的基因组区域;可以含有 5’UTR,所有 CDS,和 3’UTR
GC_signal	GC 盒;位于真核生物转录单位起始点上游的保守的富含 GC 区域,可以以多重拷贝或任一方向存在;共有序列=GGGCGG
gene	鉴定为基因的生物学意义的区域,并已经指定名称
iDNA	间插 DNA;通过几种重组中的任何一种能被消除的 DNA
intron	被转录的 DNA 区段,但通过同时剪接位于其两侧的序列(外显子)即可从转录本内部将其除去

J_segment	免疫球蛋白轻链和重链的连接区段,和 T-细胞受体 α, β 和 γ 链
LTR	长的末端重复,在确定序列的两端直接重复的序列, 类型典型地见于逆转录病毒中
mat_peptide	成熟的肽或蛋白质的编码序列;翻译后修饰之后成熟的或最终的肽或蛋白质产物的编码序列;位置不包括终止密码子(与相应的 CDS 不同)
misc_binding	不能用任何其它 Binding 关键词(primer_bind 或 protein_bind)表述的与另一个组成成分共价或非-共价结合的核酸中的位点
misc_ difference	特征序列与记载中存在的有所不同,并且不能用任何其它不同关键词(conflict, unsure, old_sequence, mutation, variation, allele 或 modified_base)表述
misc_feature	不能用任何其它的特征关键词表述的具有生物学意义的区域;新的或少见的特征
misc_recomb	任何一般性的,位点特异性的或复制的重组事件的位点,该位点中有不能用其它重组关键词(iDNA 和 virion)或来源关键词的修饰词(/transposon,/proviral)表述的双螺旋 DNA 的断裂和愈合
misc_RNA	不能用其他 RNA 关键词（prim_transcript,precursor_RNA, mRNA, 5'clip, 3'clip, 5'UTR, 3'UTR, exon, CDS, sig_peptide, transit__ __peptide, mat_peptide, intron, polyA_site, rRNA, tRNA, scRNA 和 snRNA）限定的任何转录本或 RNA 产物
misc_signal	含有控制或改变基因功能或表达之信号的任何区域,所述信号不能用其他 Signal 关键词 (promoter, CAAT_signal, TATA_signal, -35_signal, 10_signal,GC_signal,RBS,polyA_signal,enhancer,attenuator, terminator 和 rep_origin)表述
misc_structure	不能用其他 Structure 关键词(stem_loop 和 D-loop)表述的任何二级或三级结构或构象
modified_base	被指示的核苷酸是经修饰的核苷酸,并应由被指示的分子(在 mod_base 修饰词意义中给出)所取代
mRNA	信使 RNA;包括 5'非翻译区(5'UTR)，编码序列(CDS，外显子)和 3'非翻译区(3'UTR)
mutation	在此位置处，相关品系的序列中具有突然的，可遗传的变化
N_region	在重排的免疫球蛋白区段之间插入的额外的核苷酸
Old_sequence	在此位置处，所表述的序列修改了此序列以前的版本
PolyA_signal	聚腺苷酸化之后内切核酸酶裂解 RNA 转录本所必需的识别区域；共有序列＝AATAAA
PolyA_site	RNA 转录本上的位点,通过转录后聚腺苷酸化该位点将被加上腺嘌呤残基
Precursor_RNA	仍不是成熟的 RNA 产物的任何 RNA 种类；可包括 5'剪切区(5'clip)，5'非翻译区(5'UTR)，编码序列(CDS，外显子)，间插序列(内含子)，3'非翻译区(3'UTR)，和 3'剪切区(3'clip)

prim_transcript	初级（最初的，未加工的）转录本；包括5'剪切区（5'clip），5'非翻译区（5'UTR），编码序列(CDS,外显子),间插序列(内含子),3'非翻译区(3'UTR)和3'剪切区(3'clip)
prim_bind	起始复制,转录或逆转录的非一共价的引物结合位点;包括合成的例如 PCR 引物元件的位点
Promoter	参与 RNA 聚合酶的结合以启动转录的 DNA 分子区域
protein_bind	核酸上非一共价的蛋白质结合位点
RBS	核糖体结合位点
repeat_region	含有重复单位的基因组区域
repeat_unit	单个重复元件
rep_origin	复制起点;复制核酸以得到两个相同拷贝的起始位点
RRNA	成熟的核糖体 RNA；将氨基酸装配成蛋白质的核糖核蛋白颗粒（核糖体）中的 RNA 成份
S_region	免疫球蛋白重链的开关区；它参与重链 DNA 的重排,导致来自相同 B—细胞的不同免疫球蛋白类的表达
Satellite	短的基本重复单位的很多串联重复(相同或相关的);大多数具有的碱基组成或其它性质与基因组的一般水平不同,这使得它们与大部分(主带)的基因组 DNA 分离开来
ScRNA	小的细胞质 RNA;几个小的细胞质ＲＮＡ分子中的任何一个存在于真核生物的细胞质和（有时）核中
sig_peptide	信号肽编码序列；被分泌的蛋白质的Ｎ－末端结构域的编码序列；此结构域涉及新生多肽与膜的结合；前导序列
SnRNA	小的核 RNA；很多小的 RNA 种类中的任何一个都被局限于核中；几个 snRNA 参与剪接或其它 RNA 加工反应
source	鉴定序列中特定范围的生物来源；此关键词是强制性的；每一项至少要有一个跨越整个序列的单一来源关键词；每个序列可允许有一个以上的来源关键词
stem_loop	发卡结构；由 RNA 或 DNA 单链的相邻（反向）互补序列之间的碱基一配对形成的双螺旋区域
STS	序列标记位点：表述基因组上作图界标并能通过 PCR 检测的短的，单拷贝 DNA 序列；通过测定 STS 系列的次序即可作出图谱的基因组区域
TATA_signal	TATA 盒；Goldberg-Hogness 盒；在每个真核生物 RNA 聚合酶Ⅱ转录单位起点前约 25bp 处发现的保守的富含 AT 的七聚体，它可能涉及使酶定位以正确地起始；共有序列＝TATA（Ａ或Ｔ）Ａ（Ａ或Ｔ）
terminator	或者位于转录本的末端或者与启动子区域相邻的 DNA 序列，该序列可导致 RNA 聚合酶终止转录；也可以是阻抑蛋白的结合位点
transit_peptide	转运肽编码序列；核编码的细胞器蛋白质Ｎ－末端结构域的编码序列；此结构域参与将蛋白质翻译后运送到细胞器中

tRNA	成熟的转移 RNA,，小的 RNA 分子（75—85 个碱基长），介导核酸序列翻译成氨基酸序列
unsure	作者不能确定此区域的准确序列
V_region	免疫球蛋白轻链和重链的可变区，和 T 一细胞受体α，β和γ链；编码可变的氨基末端部分；可由 V__segment,D_segment,N_region 和 J_segment 组成
V_segment	免疫球蛋白轻链和重链的可变区段，和 T 一细胞受体α，β和γ链；编码大多数可变区（v_region）和前导肽的最后几个氨基酸
variation	含有来自相同基因的稳定突变的相关系列（例如 RFLP，多态性等），在此（和可能其它）位置处所述相同基因与被表述的不同
3’ clip	在加工过程中被切下的前体转录本 3’端大部分区域
3’ UTP	不被翻译成蛋白质的成熟转录本的 3’末端区域（终止密码子之后）
5’ clip	在加工过程中被切下的前体转录本 5’端大部分区域
5’ UTP	不被翻译成蛋白质的成熟转录本的 5’末端区域（起始密码子之前）
_ 10 _signal	Pribnow 盒；细菌转录单位起点上游约 10bp 处的保守区域,它可能参与结合 RNA 聚合酶;共有序列=TatAaT
_ 35 _signal	细菌转录单位起点上游约 35bp 处的保守六聚体；共有序列=TTGACa[]或 TGTTGACA[]

表 2 与蛋白质序列相关的特征关键词表

关键词	说明
CONFLICT	不同的论文报道了不同的序列
VARIANT	作者报道存在序列变体
VARSLIC	由可选择的剪接产生的序列变体的表述
MUTAGEN	经实验操作已改变的位点
MOD_RES	残基的翻译后修饰
ACETYLATION	N—末端或其它
AMIDATION	通常位于成熟的活性肽的 C—末端
BLOCKED	不能被测定的 N—或 C—末端封闭基团
FORMYLATION	N—末端甲硫氨酸的
GAMMA-CARBOXY-GLUTAMIC	天冬酰胺，天冬氨酸，脯氨酸或赖氨酸的

ACID HYDROXYLATION	
METHYLATION	通常为赖氨酸或精氨酸的
PHOSPHORYLATION	丝氨酸，苏氨酸，酪氨酸，天冬氨酸或组氨酸的
PYRROLIDONE CARBOXYLICACID	已形成内部环内酰胺的 N－末端谷氨酸
SULFATATION	通常为酪氨酸的
LIPID	脂质组成成分的共价结合
MYRISTATE	通过酰胺键与蛋白质成熟形式的 N－末端甘氨酸残基或内部的赖氨酸残基结合的豆蔻酸基团
PALMITATE	通过硫酯键与半胱氨酸残基或通过酯键与丝氨酸或苏氨酸残基结合的棕榈酸基团
FARNESYL	通过硫酯键与半胱氨酸残基结合的法尼基
GERANYL-GERANYL	通过硫酯键与半胱氨酸残基结合的香叶基－香叶基基团
GPI__ANCHOR	与蛋白质成熟形式 C －末端残基的 α －羧基相连的糖基－磷脂酰肌醇（GPI）基团
N__ACYL DIGLYCERIDE	原核生物脂蛋白成熟形式的 N－末端半胱氨酸，所述脂蛋白具有酰胺－键联的脂肪酸和通过酯键连接了两个脂肪酸的甘油基
DISULFID	二硫键；“FROM”和“TO”终点表示通过一个链－内二硫键连接的两个残基；如果“FROM”和“TO”终点是完全相同的，则二硫键是链－间键，而说明书领域示出交联的性质
THIOLEST	硫醇酯键；“FROM”和“TO”终点表示通过硫醇酯键连接的两个残基
THIOETH	硫醚键；“FROM”和“TO”终点表示通过硫醚键连接的两个残基
CARBOHYD	糖基化位点；碳水化物（如果已知）的性质在说明书领域给出
METAL	金属离子的结合位点；说明书领域示出金属的性质
BINDING	任何化学基团（辅酶，辅基，等等）的结合位点；基团的化学性质在说明书领域给出
SIGNAL	信号序列的范围（前肽）
TRANSIT	运转肽的范围（线粒体，叶绿体或微体）
PROPEP	前肽的范围
CHAIN	成熟蛋白质中多肽链的范围
PEPTIDE	被释放的活性肽的范围

DOMAIN	序列中感兴趣的区域的范围；所述区域的特征在说明书领域给出
CA__BIND	钙－结合区域的范围
DNA__BIND	DNA-－结合区域的范围
NP_BIND	核苷酸磷酸酯结合区域；核苷酸磷酸酯的特征示于说明书领域
TRANSMEM	转膜区域的范围
ZN_FING	锌指区域的范围
SIMILAR	与另一个蛋白质序列具有相似性的区域；与那个序列有关的精确的资料在说明书领域给出
REPEAT	内部序列重复的范围
HELIX	二级结构；螺旋，例如α－螺旋，３（10）螺旋，或Pi-螺旋
STRAND	二级结构；β－链，例如氢键连接的β－链，或分离的β－桥中的残基
TURN	二级结构转角，例如Ｈ－键连的转角（３－转角，４－转角或５－转角）
ACT_SITE	涉及酶活性的氨基酸
SITE	序列中任何其它感兴趣的位点
INIT_MET	已知序列以起始密码子甲硫氨酸开始
NON_TER	序列末端的残基不是末端残基；如果应用于位置１，这表示第一个位置不是完整分子的Ｎ－末端；如果应用于最后一个位置，这表示此位置不是完整分子的Ｃ－末端；对此关键词没有说明书领域
NON_CONS	非连串残基；表示序列中的两个残基不是连串的，在它们之间有很多末测序的残基
UNSURE	序列的不确定性；用于表述不能确定序列排列的序列区域